Universidad Carlos III de Madrid
Escuela Politécnica Superior
Ingeniería en Informática

Proyecto Fin de Carrera

ANÁLISIS Y DISEÑO DE UN DATA MART PARA EL SEGUIMIENTO ACADÉMICO DE ALUMNOS EN UN ENTORNO UNIVERSITARIO

Autor: Miguel Rodríguez Sanz
Tutor: Jack Mario Mingo Postiglioni
Fecha: 22 Julio, 2010

Agradecimientos

En primer lugar a mi tutor Jack Mario, por su gran predisposición durante todos los meses en los que he estado trabajando en el proyecto y haberme dedicado gran parte de su tiempo para que pudiera concluirlo.

A mis padres Antonio y Paloma, por animarme todo este tiempo a lograr mis metas y llegar hasta donde estoy hoy.

A Luz, por sacrificar fines de semana y vacaciones para que pudiera avanzar en el proyecto y pudiera concluirlo echándome una mano en todo lo que podía.

Y por último a mis compañeros y ahora amigos de la universidad, con alguno de los cuáles comparto vida laboral actualmente. Víctor G., Víctor R., Mikel, David, Igor, José Ignacio y Alfredo, muchas gracias.

"No pretendamos que las cosas cambien, si siempre hacemos lo mismo. La crisis es la mejor bendición que puede sucederle a personas y países, porque la crisis trae progresos.

La creatividad nace de la angustia como el día nace de la noche oscura. Es en la crisis que nace la inventiva, los descubrimientos y las grandes estrategias. Quien supera la crisis se supera a sí mismo sin quedar superado. Quien atribuye a la crisis sus fracasos y penurias, violenta su propio talento y respeta más a los problemas que a las soluciones.

La verdadera crisis, es la crisis de la incompetencia.

El inconveniente de las personas y los países es la pereza para encontrar las salidas y soluciones. Sin crisis no hay desafíos, sin desafíos la vida es una rutina, una lenta agonía. Sin crisis no hay méritos. Es en la crisis donde aflora lo mejor de cada uno, porque sin crisis todo viento es caricia.

Hablar de crisis es promoverla, y callar en la crisis es exaltar el conformismo.

En vez de esto, trabajemos duro. Acabemos de una vez con la única crisis amenazadora, que es la tragedia de no querer luchar por superarla"

Albert Einstein (1879 – 1955)

Índice

Lista de figuras

Lista de tablas

Acrónimos y definiciones

Back-end	Término que identifica el comienzo de un proceso.
Batch	Ciclo donde se procesan muchos registros uno tras otro sin que intervenga interactivamente el usuario.
BBDD	Bases de datos
BDL	Business Dimensional Lifecycle.
BDM	Business dimensional model (MDN).
BI	Business Intelligence
CSAE	Consejo superior de administración electrónica.
Cubo	Es una base de datos multidimensional en la cual el almacenamiento físico de los datos se realiza en forma de vector multidimensional.
DBMS	Database Management System
DWH	Data Warehouse
DM	Data Mart.
DSS	Sistemas de Soporte de Decisiones. (Decision Support Systems).
EIS	Executive Information Systems.
ERP	Sistemas de Planificación de Recursos Empresariales (Enterprise Resource Planning, ERP).
ETL	ETL son las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load).
Front-end	Término que identifica el final de un proceso.
GUI	Graphical User Interface
ID	Identificador.
Join	Sentencia SQL que permite combinar registros de dos o más tablas en una base de datos.
MDN	Modelo Dimensional del Negocio (MDN).
Metadatos	Los metadatos son datos altamente estructurados que describen información, describen el contenido, la calidad, la condición y otras características de los datos.
MOLAP	Procesamiento analítico en línea multidimensional.
NIA	Número de identificación del alumno.
OLAP	OnLine Analytical Processing o Procesamiento Analítico En Línea
Raw data	*Datos que aún no han sido procesados o analizados.*
ROLAP	Procesamiento analítico en línea relacional.
RWM	Rapid Warehousing Methodology (RWM).
Snowflake	Término anglosajón para el modelo de copo de nieve.
Sponsor	Personas, instituciones o departamentos que promueven la implantación del Data Warehouse.
Summary Data	Datos resultantes de un proceso de transformación
Vector Multidimensional	Es un vector que se indexa mediante una lista ordenada de enteros.

1. Introducción

Hoy en día, las bases de datos (BBDD) existentes en las empresas mantienen la información necesaria para la actividad diaria de la organización, ya que dichas BBDD suministran datos a los sistemas de información corporativos. Éstas, representan una herramienta imprescindible en el mundo actual, aunque no suficiente para el correcto funcionamiento de los sistemas de información de cualquier organización. Es importante por ello desde el punto de vista de la organización empresarial, que la estrategia en el área de sistemas y tecnologías de la información vaya encaminada hacia una correcta política de mantenimiento, actualización y gestión de las BBDD y una adecuada política de I+D+I en la organización.

En este sentido, y puesto que los cambios que se producen actualmente en las tecnologías y sistemas de información son demasiado rápidos, en este proyecto hacemos un análisis y diseño de una herramienta llamada Data Warehouse. Esta tecnología de la información representa el último avance dentro de las bases de datos, y se configura como el entorno idóneo para la consulta y el análisis de la información procedente tanto de los sistemas transaccionales internos, como de las fuentes de información externas de interés para la empresa.

La finalidad del Data Warehouse consiste en convertir los datos contenidos en las bases de datos corporativas de las organizaciones, en información y ésta, a su vez, en conocimiento útil en el proceso de toma de decisiones estratégicas. El Data Warehouse es una herramienta que va a permitir a los directivos de las organizaciones formular preguntas, realizar consultas y analizar los datos en el momento, forma y cantidad que precisen sin necesidad de tener que acudir al personal informático de la empresa.

Desde mediados de los años ochenta, en los que las tecnologías de la información se esforzaban por automatizar los procesos de tipo repetitivo o administrativo haciendo uso de los sistemas de información operacionales, los Data Warehouse han sufrido una gran evolución. En los últimos años, el concepto de Data Warehouse ha ido perfeccionándose (gracias al aumento de la capacidad de almacenamiento, la expansión de internet y las nuevas herramientas de consulta de datos) y adaptándose a las necesidades crecientes de información en las empresas de forma que los actuales Data Warehouse pueden proporcionar soluciones a todo tipo de usuarios.

Por último, cabe mencionar la existencia del *Data Mart* que por ahora podríamos decir que son una versión más reducida de un Data Warehouse. Estos *Data Mart* a menudo contienen información específica de algún departamento concreto de la organización como pueden ser marketing o finanzas. Idealmente, estos *Data Mart* deberían ser un subconjunto del Data Warehouse, a fin de mantener consistencia de datos corporativos y mantener la seguridad e integridad de la información que se está usando.

Hoy en día, debido al coste de desarrollo e implantación de un Data Warehouse en la organización, se hace patente una mayor demanda de *Data Mart* en las mismas. Como veremos más adelante elegir una u otra opción ofrecerá una serie de ventajas y desventajas para la empresa que quiera poner en marcha su implantación. Como vemos

en la figura 1 obtenida del CSAE, el uso previsto de los *Data Mart* en el futuro crece mientras que el de los Data Warehouse disminuye.

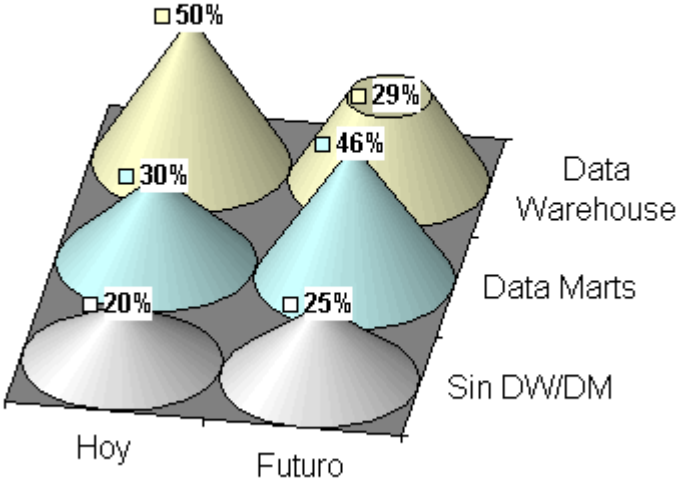

Figura 1: Data Warehousing en la actualidad y en el futuro.

Es importante saber que el Data Warehousing es una tecnología que se está implantando en numerosos ámbitos empresariales. En la figura 2 se puede observar cuáles son los principales sistemas empresariales en los que se utiliza esta tecnología.

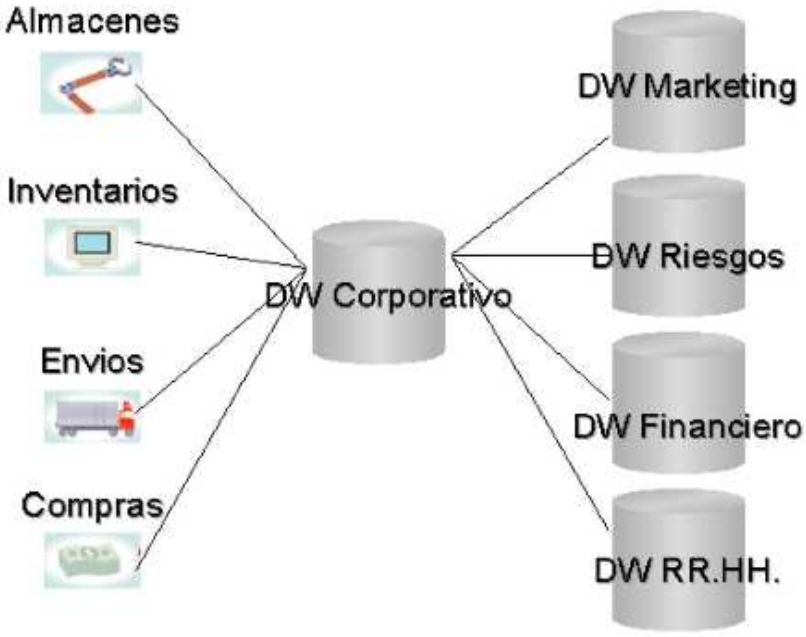

Figura 2: Principales departamentos con Data Warehousing.

En los siguientes apartados del documento se dará una visión clara y concisa del origen y los fundamentos de los Data Warehouse y los Data Mart desde sus comienzos, y el propósito para el que hoy en día se construyen, principalmente en el ámbito empresarial. Se propondrá una metodología clara y estructurada para poder desarrollar

un Data Mart aplicado a un entorno de seguimiento académico de alumnos universitarios.

Se puede decir que un Data Warehouse en el entorno universitario posibilitará al usuario, cualquiera de las universidades del territorio español, la obtención de datos precisos acerca de las carreras universitarias, el alumnado, las asignaturas y la calificación de las mismas con el fin de obtener un mejor rendimiento en su conjunto y poder tomar decisiones que aumenten las posibilidades de éxito de los alumnos en los casos en que sea necesario.

El presente documento se estructura de la siguiente forma: En la primera parte se hace un repaso por los entornos (capítulo dos), metodologías (capítulo tres), modelos y arquitecturas (capítulo cuatro) principalmente empleados en la construcción de los Data Warehouse, obteniendo un sólido fundamento teórico sobre el que construir el desarrollo que constituye el propósito del proyecto. En la segunda parte se describe con detalle el desarrollo de un *Data Mart* para seguimiento académico de alumnos, dedicando un capítulo a cada una de las fases de la metodología empleada para su construcción. Así, el capítulo uno de este bloque describe el plan de proyecto, el capítulo dos la definición de los requerimientos del negocio, el capítulo tres el modelado dimensional, el capítulo cuatro el diseño físico, el capítulo cinco el desarrollo y presentación de los datos, el capítulo seis la arquitectura técnica, el capítulo siete la selección de los productos y su instalación, el capítulo ocho la especificación de aplicaciones para usuarios finales, el capítulo nueve el desarrollo de aplicaciones para usuarios finales, el capítulo diez la implementación, el capítulo once las cuestiones relativas al mantenimiento y crecimiento del *Data Mart* en cuestión y, por último el capítulo doce describe la gestión del proyecto. Para finalizar se comentan las conclusiones y las futuras líneas de trabajo que se han detectado en la realización de este proyecto.

2. Objetivos del proyecto

El propósito de la realización de este proyecto es la implantación de un Data Mart realizando las fases y actividades propuestas por la metodología que mejor se adapte al objeto de negocio concreto. El presente documento se centrará en el planteamiento, solución y desarrollo de dicha tecnología dentro del proceso educativo universitario con el fin de conocer la respuesta de los diferentes alumnos a los estudios que han comenzado.

Los objetivos del proyecto son dobles: en una primera parte se presentará un repaso de los conceptos fundamentales relacionados con los Data Warehouse y los Data Mart, haciendo especial hincapié en las metodologías, modelos y arquitecturas empleados principalmente en estas tecnologías; mientras que en una segunda parte se aplicará una metodología concreta al desarrollo de un Data Mart que permita responder a las necesidades de negocio del área académica de una universidad, específicamente en relación al seguimiento de los alumnos en las diferentes titulaciones ofertadas. De esta forma, se puede considerar que la primera parte actúa como base teórica para guiar las decisiones tomadas en la segunda parte. Resumiendo, los objetivos del proyecto son:

1. Realizar un estudio preliminar que permita:
   - Revisar los conceptos fundamentales asociados con la tecnología de los Data Warehouse y los Data Marts.
   - Analizar las propuestas metodológicas más habituales para el desarrollo de Data Warehouse y Data Marts.
   - Establecer un marco de trabajo que permita orientar la implementación práctica de proyectos de Data Warehouse y Data Marts.

2. Aplicar una metodología apropiada a un proyecto de seguimiento académico de alumnos en el entorno universitario, con especial énfasis en las fases de diseño de la arquitectura y el modelado de datos.

Por último es importante destacar que dentro de los objetivos del proyecto no se contempla la realización de la parte visual y de interacción de los usuarios con el sistema que corresponde con las aplicaciones de usuario final. De esta manera los puntos del proyecto que se refieren al desarrollo de éstas y a la implantación real del sistema se limitan a ser ligeras recomendaciones propias de la metodología seleccionada, ya que estos apartados exceden los límites del trabajo planteados al comienzo del mismo.

Parte I: Fundamentos de los Data Warehouse y los Data Mart

1. Panorama actual de los Data Warehouse y los Data Mart

Desde hace varias décadas las organizaciones empresariales han buscado en el almacenamiento de datos de sus sistemas operacionales soluciones que les ayuden a atender sus necesidades a la hora de tomar decisiones de negocio.

Algunas de estas organizaciones han extraído los datos de sus bases de datos para combinarlos de varias formas estructuradas o no estructuradas, en su intento por atender a los usuarios en sus necesidades de información, mientras que otras permiten un acceso directo a la información que hay recogida dentro de sus bases de datos operacionales. En cualquiera de los dos casos se ha podido observar una gran evolución en el almacenamiento de estos datos, sirviéndose de ellos para las importantes decisiones que debiera tomar la empresa.

Actualmente, las **bases de datos operacionales** son útiles en un entorno muy concreto que responde a las necesidades para las que se crearon. Estas necesidades suelen involucrar entornos de gestión puros en los que las características principales de las operaciones suelen ser las de la simplicidad en las consultas y tipos de los datos.

Por otra parte, las necesidades de información hoy en día han variado. La disponibilidad de gran cantidad de información es de vital importancia para los negocios, ya que las decisiones de futuro se suelen tomar sobre la base de dicha información. Este tipo de necesidades para reflejar tendencias, evoluciones, hechos históricos en el negocio y posibilidades futuras son algo que la alta dirección de las instituciones o empresas debe manejar y maneja de una forma habitual y es la causante de que hayan aparecido en el mercado **herramientas denominadas "ayudas a la toma de decisiones".**

Los procesos que obtienen la información útil para la toma de decisiones con la estructura adecuada son procesos que, a priori, **involucran un alto coste en consumo de recursos** dado el gran volumen de datos sobre el que actúan y el tiempo de procesamiento que conlleva la obtención de las nuevas consultas.

Está claro, por lo dicho antes, que las denominadas "bases de datos operacionales" de la empresa no se pueden ver afectadas en sus tiempos de respuesta por dicho consumo en recursos. Esto, unido al hecho de que la estructura de las bases de datos de producción puede no ser la óptima para la obtención de la información deseada por la dirección de la empresa, obliga a la aparición de una nueva forma de almacenar dicha información: el **Data Warehouse (DWH),** que consiste en una réplica masiva de los datos disponibles en las bases de datos operacionales, de tal forma que su estructura ya no responda a las necesidades del modelo relacional puro, que suele emplearse en los sistemas operacionales, puesto que sobre ella no se van a efectuar el mismo tipo de operaciones que se hacen sobre los citados sistemas operacionales.

Por lo dicho hasta ahora parece que se ha definido un DWH como una simple copia de los datos sobre la que efectuar las consultas de negocio de la empresa. Sin embargo, la problemática asociada a la obtención y tratamiento que de sus datos se hace, convierten al DWH en una nueva estructura con una problemática asociada muy concreta y diferenciada de las bases de datos operacionales.

Desde un punto de vista técnico, un DWH es un almacén de datos nuevos y antiguos, con ciertas propiedades (no volátil, variable en el tiempo, integrado [1]) con el que la empresa puede realizar un profundo análisis de su organización y apoyar sus decisiones de negocio sobre las consultas realizadas a partir de dichos datos. Este almacén de datos contiene toda la información relativa a una misma temática de la empresa con el fin de agrupar los datos y unirlos entre sí.

Basándonos en la definición que nos proporciona Bill Inmon [1], podemos definir un DWH como un repositorio de datos con las siguientes propiedades:

- **Orientado a temas -** Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- **Variante en el tiempo -** Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.
- **No volátil.-** La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura y se mantiene para futuras consultas.
- **Integrado.-** La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes

En base a esta definición se puede ver claramente que la existencia de un DWH no conlleva exclusivamente el hecho de que se realice una copia masiva de datos, sino que esa copia tiene un determinado fin y que su propia existencia involucra una dinámica de trabajo diferenciada.

Inicialmente, la primera diferencia entre ambas formas de almacenamiento estriba en la orientación de los datos y su organización: en el mundo operacional, los datos están orientados a los procesos o transacciones, mientras que en el entorno de toma de decisiones los datos deben estar orientados al tema, entendiendo por tal los conceptos que tengan relevancia en la organización en la que existe. Por sí solo, este hecho ya implica la existencia de procesos diferentes a los existentes en las bases de datos operacionales. Así, parece evidente que uno de los procesos característicos de un entorno de DWH sea el proceso de "llenado", ya que la información del DWH proviene de los datos existentes en las bases de datos operacionales de la organización. Este proceso no debe afectar de ninguna manera a los rendimientos de dichos sistemas, con lo que se deben buscar los momentos adecuados en los que realizarlo, lo que conlleva modificaciones en la "mecánica de trabajo" de la organización.

Otra de las características de un DWH reside en que el diseño debe orientarse a las nuevas necesidades de tiempo de vida útil mínimo de la base de datos y a que los datos son no volátiles, lo que hace de **la fecha un atributo importantísimo** a la hora de convertir en "historiados" a dichos datos. El usuario del DWH no efectúa operaciones de escritura sobre él, ya que solamente se carga con el proceso de llenado ya

mencionado. Este proceso de carga es el que permite al DWH mantener la propiedad de variación en el tiempo.

Por otra parte y debido a la "no volatilidad", es de destacar el rápido crecimiento en volumen de datos, lo que obliga al diseño y mantenimiento de políticas de almacenamiento y de realización de **copias de seguridad que sean eficientes**.

Por último, hay que decir que el acceso al DWH es diferente al tradicional, y que, otra vez debido a ese gran volumen, se debe dedicar un gran esfuerzo a optimizar el acceso a los datos por parte de las consultas. Pequeños fallos de diseño en las consultas adquieren ahora gran relevancia dado que se notará mucho la falta de optimización.

Todo esto nos lleva a poder representar el flujo de información existente en un DWH como un flujo lineal que comienza en las bases de datos tradicionales y que tiene como característica que la escritura o actualización se produce en un punto muy concreto y que el otro extremo solamente efectúa operaciones de lectura.

Todas estas características planteadas por Bill Inmon [1], se engloban en una metodología **top-down**, según la cual, debemos ir desde una visión más general de las distintas partes que componen nuestro almacén y posteriormente ir concretando y refinando cada una de las partes por separado. Por ello, según este autor, una vez que hemos desarrollado nuestro DWH, es cuando podemos abordar el desarrollo de los **Data Mart (DM)**. Los DM son subconjuntos de datos de un DWH para áreas específicas. Entre las características de un Data Mart destacan:

- Usuarios limitados.
- Área específica.
- Tiene un propósito específico.
- Tiene una función de apoyo

Sin embargo, esta metodología, se contrapone con la metodología **bottom-up** que defienden otros autores como Ralph Kimball [2], el cual define un DWH como:

"Una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis".

Según el mismo Kimball, un DWH no es más que: "la unión de todos los Data Marts de una entidad". Ahora bien, una vez almacenados los datos de la empresa, se pueden emplear aplicaciones para la obtención estructurada de lo que se quiera consultar en cada momento.

Sin embargo y como afirman los dos autores, los DWH se diferencian en muchos factores con respecto a los entornos operacionales, lo que les hace tener una mayor potencia a la hora de realizar búsquedas sobre los datos en entornos orientados a la toma de decisión, aunque en otro tipo de situaciones las BBDD operacionales podrían funcionar mejor. Las principales diferencias entre ambos sistemas se podrían resumir en:

- Los DWH, normalmente emplean esquemas desnormalizados para optimizar el desarrollo de las consultas mientras que los entornos operacionales emplean esquemas completamente normalizados para optimizar el desarrollo de

actualizaciones, inserciones y borrados y así, garantizar la consistencia de los datos.

- Los entornos operacionales están orientados a una serie de operaciones predefinidas, mientras que los DWH son diseñados para realizar consultas más genéricas. Esto es debido a que no se puede conocer de antemano la carga de trabajo del mismo, por lo que debe ser optimizado para funcionar bien para una gran variedad de operaciones de consulta.

- En cuanto a modificaciones, los DWH se actualizan normalmente mediante procesos de extracción, transformación y carga de datos (Extraction, Transformation and Load, ETL) utilizando técnicas de modificación de datos a gran escala, no siendo los usuarios finales quienes realizan las actualizaciones. Por otro lado, los entornos operacionales, siempre contemplan los datos hasta la fecha en curso y reflejan el estado actual de cada transacción.

- Un entorno operacional normalmente tiene unos datos históricos que comprenden unas pocas semanas, ya que lo que buscan es contener solamente los datos necesarios para la operativa cotidiana de la organización. Por otro lado, los DWH contienen datos de varios meses y años con el fin de realizar análisis históricos. Así, una consulta común en un DWH puede obtener miles de filas, mientras que en un entorno operacional los resultados obtenidos tendrán un valor muchísimo más reducido.

Con todo esto, las aplicaciones para soporte de decisiones basadas en un DWH pueden hacer más práctica y fácil la explotación de los datos para una mayor eficacia del negocio, objetivo que no se logra cuando se usan sólo los datos que provienen de las aplicaciones operacionales puesto que resulta ser complejo y menos eficaz.

Por todos los aspectos descritos, se puede decir que las propiedades de los DWH los hacen especialmente idóneos para estudiar el comportamiento del alumnado universitario a lo largo del proceso educativo ya que permiten su análisis desde diferentes perspectivas, así como valorar los resultados académicos del mismo. Este análisis permitirá obtener información acerca de qué carreras son más complejas mediante porcentajes de aprobados, tasas de abandono universitario, fechas de finalización de las carreras por parte de los alumnos y multitud de datos relevantes acerca del proceso educativo en estudio.

2. Entorno de los Data Warehouse

Cada día, el volumen de datos que se maneja en las empresas a partir de los sistemas de planificación de recursos empresariales (Enterprise Resource Planning, ERP) aumenta exponencialmente. Si además de seguir acumulando datos de todas aquellas funciones y departamentos ya automatizados, se siguen automatizando procesos, y con todo ello no es suficiente para obtener la información necesaria en la toma de las decisiones empresariales, se hace necesario el desarrollo de un DWH. Cuando una empresa se embarca en su desarrollo, busca [3]:

- *Mayor poder de procesamiento y sofisticación de herramientas*
- *Demanda de mejora del acceso a datos*
- *Necesidad de información para la toma de decisiones*
- *Recopilación de información* $\Rightarrow$ *Alto Coste*

Con todo ello, cuando una empresa quiere abordar un proyecto de DWH se enfrenta con ciertos aspectos que es necesario analizar antes de su desarrollo, con el fin de decidir instalar o no esta tecnología en la misma [3].

2.1. Motivos para desarrollar un Data Warehouse

Hay muchas ventajas por las que es recomendable usar un almacén de datos y por las que una empresa debe decidirse a implantarlo. Entre ellas podamos citar las siguientes:

- Los almacenes de datos facilitan el funcionamiento de las aplicaciones de los sistemas de apoyo a la decisión tales como *informes de tendencia*, por ejemplo: obtener los ítems con la mayoría de las ventas en un área en particular dentro de los últimos dos años; *informes de excepción*, informes que muestran los resultados reales frente a los objetivos planteados a priori.

- Los almacenes de datos pueden trabajar de manera conjunta para así aumentar el valor operacional de las aplicaciones empresariales, en especial la gestión de relaciones con clientes conocidas como *Customer Relationship Management* (CRM).

- Acceso a toda la información de la empresa. La información que proviene de sistemas de origen diferentes se consolida, sin importar si provienen de la misma o varias fuentes.

- Consistencia de la información al consolidarla desde varios departamentos origen a un solo destino. Esto facilita la posterior toma de decisiones al poder hacer un mejor análisis de la información.

- Beneficios en costes, tiempos y productividad de la empresa. Un DWH ayuda a obtener mejores tiempos de respuesta y supone una mejora en los procesos de producción.

Por todo ello se dice que si una empresa quiere eficacia en los negocios que le competen, tomar decisiones cercanas a sus clientes y una ventaja competitiva, lo ideal es implementar un DWH que le ayude a obtener esos beneficios.

Sin embargo, utilizar almacenes de datos también plantea algunos inconvenientes, como por ejemplo:

- Una herramienta tan imprescindible para la empresa de hoy, necesita de un constante y costoso soporte técnico. Debido a su complejidad, el DWH es muy "frágil" en cuanto a su funcionamiento se refiere, por lo que se hace necesaria su continua revisión.

- Los almacenes de datos se pueden quedar obsoletos relativamente pronto.

- Una vez implementado puede ser complicado añadir nuevas fuentes de datos.

- En un proceso de implantación pueden encontrarse dificultades ante los diferentes objetivos que pretende una organización.

- El diseño e implementación del DWH resulta caro (aunque puede considerarse como una inversión) y toma mucho tiempo puesto que se necesitan obtener muchos datos, y estos se tienen que organizar de la mejor manera para que el DWH pueda realmente cumplir con su tarea.

Después de valorar las ventajas y los inconvenientes la empresa debería analizar la decisión de si es conveniente o no su desarrollo.

2.1.1. Situación de partida

Cualquier solución propuesta de DWH debe estar determinada por las necesidades del negocio en el que se desarrolle y debe ser compatible con la arquitectura técnica existente y planificada de la compañía. En caso de no encajar en la arquitectura existente habría que valorar si los beneficios a obtener, compensan la fuerte inversión a realizar en nuevas arquitecturas.

Como punto de partida nos podríamos apoyar en una expresión muy extendida en el mundo del Data Warehousing que algunos autores afirman:

"Un Data Warehouse no se puede comprar, se tiene que construir".

Esto quiere decir que la construcción e implantación de un DWH es un proceso evolutivo que se tiene que apoyar en una metodología específica para este tipo de procesos, si bien es más importante que la elección de la mejor de las metodologías, el realizar un control para asegurar el seguimiento de la misma.

En las fases que se establezcan en el alcance del proyecto es fundamental el incluir una fase de formación en la herramienta utilizada para un máximo

aprovechamiento de la aplicación. Seguir los pasos de la metodología y comenzar el DWH por un área específica de la empresa, nos permitirá obtener resultados tangibles en un corto espacio de tiempo.

2.1.2. Tipo y características del negocio

Es indispensable tener el conocimiento exacto sobre el tipo de negocios de la organización y el soporte que representa la información dentro de todo su proceso de toma de decisiones.

Antes de implantar una herramienta tecnológica, es necesario que la empresa cree una cultura de administración del conocimiento, en donde se haga notar la importancia de compartir la información, de tal manera que no solo una persona conozca el funcionamiento del negocio.

La función del DWH es contribuir a identificar elementos desconocidos del entorno de negocio y transformar la información en conocimiento a partir del análisis detallado de los datos recopilados. Es importante no confundir una gran base de datos estructurada con un auténtico DWH ya que éste trata de crear un método, un sistema para explotar auténticamente la información en beneficio del negocio, donde lo importante es la aplicación, el proceso, la manera en que se aplica el conocimiento del negocio.

2.1.3. Entorno técnico

Se deben incluir aspectos tanto del hardware como de las aplicaciones y herramientas haciendo especial énfasis en los Sistemas de Soporte de Decisiones (DSS). Estos DSS son herramientas de *Business Intelligence (BI)* enfocadas al análisis de los datos de una organización.

En principio, puede parecer que el análisis de datos es un proceso sencillo y fácil de conseguir mediante una aplicación hecha a medida o un ERP sofisticado. Sin embargo, no es así, estas aplicaciones suelen disponer de una serie de informes predefinidos en los que presentan la información de manera estática, pero no permiten profundizar en los datos o navegar entre ellos.

El DSS es una de las herramientas más emblemáticas del BI ya que, entre otras propiedades, permiten resolver gran parte de las limitaciones de los programas de gestión.

2.1.4. Expectativas de los usuarios

Un proyecto de DWH no es únicamente un proyecto tecnológico, es una forma de vida de las organizaciones y como tal, tiene que contar con el apoyo de todo el personal, tanto de los usuarios como de la alta dirección de la empresa. Además, como ya se comentó en el apartado sobre las características del negocio, se ha de resaltar la importancia de compartir la información, de tal manera que no solo una persona conozca el funcionamiento del negocio.

2.1.5. Etapas de desarrollo

A partir del conocimiento acumulado en los apartados anteriores se puede comenzar el desarrollo de un modelo conceptual para la construcción del DWH. Para llevar a cabo este desarrollo podemos distinguir en dicha construcción dos etapas principales: la definición de una arquitectura DWH y la construcción de los DM de manera incremental. Más adelante haremos hincapié en la arquitectura del DWH.

Además, es necesario el desarrollo de un prototipo destinado a simular tanto como sea posible el producto final que será entregado a los usuarios, así como un proyecto piloto de DWH, que será cada uno de los resultados generados de forma iterativa para llegar a la construcción del producto final deseado.

3. Metodologías de desarrollo de un Data Warehouse

El desarrollo de un DWH debe tener en cuenta las necesidades de los usuarios en cuento a la presentación de informes y análisis. De otro modo, el almacén de datos se convertirá en un cajón de datos del que será difícil extraer la información que los usuarios necesitan.

Para que un DWH pueda conseguir su objetivo, los procesos de negocio se seleccionan con el objetivo de modelarlos, estableciendo una granularidad para cada uno de ellos. Por este motivo es muy importante entender correctamente los datos de los diferentes sistemas dentro de la organización y las relaciones entre ellos. La gestión de estas relaciones durante la carga de almacenamiento de datos es esencial.

En cuanto a su desarrollo, a la hora de abordar un DWH no hay una única metodología en la que basar el diseño, sino que dependiendo del contexto en el que se encuentre la empresa y los objetivos que persiga se puede emplear una u otra metodología. Estas diferentes metodologías se pueden englobar dentro de dos grandes bloques: **top-down** y **bottom-up** que se corresponden con las metodologías propuestas por Bill Inmon y Ralph Kimball respectivamente. Estos autores merecen una especial atención porque, en muchos aspectos, se consideran los precursores del DWH y sus opiniones son muy valoradas en la industria.

El enfoque **top-down** se utiliza cuando la tecnología y los problemas del negocio se conocen de antemano. Este enfoque logra la sinergia entre los problemas de negocio alcanzando los objetivos buscados. Se trata de un método sistémico, que minimiza los problemas de integración, pero es costoso, debido a la gran cantidad de datos y su poca flexibilidad. En este método se formula un resumen del sistema, sin especificar detalles. Cada parte del sistema se refina diseñándola con mayor detalle. Después, cada parte nueva se redefine, cada vez con mayor detalle, hasta que la especificación completa es lo suficientemente detallada para validar el modelo. Este modelo se diseña con frecuencia con la ayuda de "cajas negras" que hacen más fácil cumplir requerimientos, aunque estas cajas negras no expliquen en detalle los componentes individuales.

El enfoque **top-down** se adapta a la visión de Bill Inmon [1], quien considera que el almacén de datos debe responder a las necesidades de todos los usuarios en la organización, y no sólo de un determinado grupo.

Por otro lado el enfoque **bottom-up** es una metodología rápida que se basa en experimentos y prototipos. Es un método flexible que permite a la organización ir más lejos con menores costos. La idea es construir DM independientes para evaluar las ventajas del nuevo sistema a medida que avanzamos. En él, las partes individuales se diseñan con detalle y luego se enlazan para formar componentes más grandes, que a su vez se enlazan hasta que se forma el sistema completo. Las estrategias basadas en el flujo de información **bottom-up** se antojan potencialmente necesarias y suficientes porque se basan en el conocimiento de todas las variables que pueden afectar a los elementos del sistema.

Sin embargo, podría haber problemas tratando de integrar los DM en un DWH empresarial ya que la primera iteración de definición de datos y las siguientes puede que no sean compatibles.

Este enfoque se adapta a la visión de Ralph Kimball [2], que considera que el almacén de datos tiene que ser entendido fácilmente por los usuarios y ofrecer respuestas correctas a la mayor brevedad posible. Este enfoque parte de los requisitos de negocio, mientras que el enfoque **top-down** propone la validación de los requisitos una vez que se tiene el sistema.

3.1. Metodologías más empleadas dentro del modelo top-down

3.1.1. Metodología propuesta por Bill Inmon

Esta metodología la definió su autor en el año 1992 en el libro "*Building the Data Warehouse"* [1]. En él proponía los mecanismos necesarios para llevar a cabo la correcta realización de un DWH.

Para Bill Inmon, el diseño de un DWH comienza ya con la mera introducción de datos en el mismo, debido a las grandes cargas de datos que deben hacerse antes de su introducción en el DWH, dependiendo de ello la eficiencia de estos sistemas para acceder a los datos. Además, la definición de Inmon sustenta uno de los principios fundamentales del desarrollo de un DWH, el principio que el ambiente de origen de los datos y el ambiente de acceso de datos deben estar físicamente separados en diferentes bases de datos y en equipos separados. Por último, los actuales sistemas tienen gran cantidad de datos, lo que hace poco realista el intentar hacer cargas cada poco tiempo. Si el volumen de datos no está cuidadosamente gestionado y condensado, dicho volumen de datos impide que los objetivos del DWH se alcancen.

A Inmon se le asocia frecuentemente con los DWH a nivel empresarial, que involucran desde un inicio todo el ámbito corporativo, sin centrarse en un incremento específico hasta después de haber terminado completamente el diseño del DWH. En su filosofía, un DM es sólo una de las capas del DWH y los DM son dependientes del depósito central de datos o DWH Corporativo y por lo tanto se construyen después de él. El enfoque de Inmon de desarrollar una estrategia de DWH e identificar las áreas principales desde el inicio del proyecto es necesario para asegurar una solución integral ya que esto ayuda a evitar la aparición de situaciones inesperadas que puedan poner en peligro el proyecto, debido a que se conoce con antelación y bastante exactitud la estructura que presentarán los principales núcleos del desarrollo, lo que permite enfocar los esfuerzos del desarrollo actual para ser compatible con los subsiguientes.

Inmon es defensor de utilizar el modelo relacional para el ambiente en el que se implementará el DWH Corporativo, ya que como él mismo afirma, la creación de una base de datos relacional con una ligera normalización, son la base de los DM. O lo que es lo mismo, a partir de los esquemas relacionales, a los que se les irá añadiendo complejidad, se obtendrán finalmente los DM.

El desarrollo de la metodología propuesta por Inmon en [1] se aprecia en la siguiente figura:

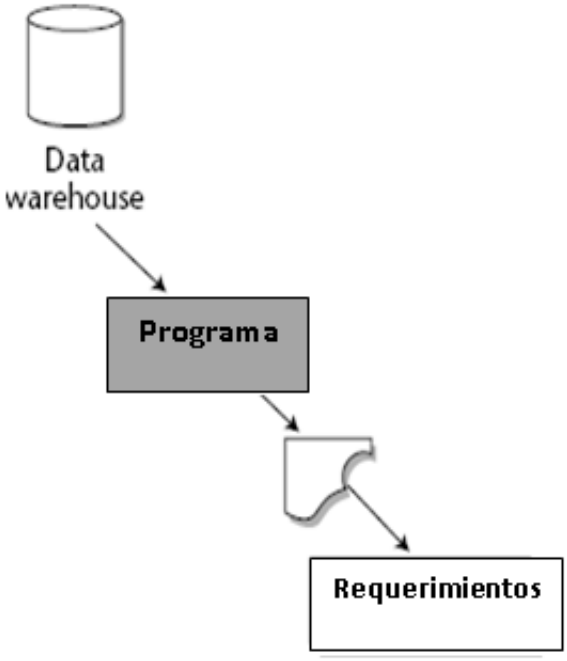

Figura 3: Desarrollo del DWH según la metodología de Bill Inmon

La metodología de Inmon tiene un enfoque a modo de explosión en el sentido de que en cierto modo no viene acompañada del ciclo de vida normal de las aplicaciones, sino que los requisitos irán acompañando al proyecto según vaya comprobándose su necesidad. Esta visión de Inmon puede traer consigo mucho riesgo a la compañía, ya que invierte grandes esfuerzos en el desarrollo del DW y no es hasta la aparición de los DM cuando se empieza a explotar la inversión y obtener beneficios.

Esta estrategia se contempla en el marco de que es imposible conocer cuáles son las necesidades concretas de información de una empresa, el ambiente dinámico en que se mueve la organización, el cambio de estructura que conlleva el desarrollo de la nueva plataforma y los consiguientes cambios a los sistemas transaccionales que su introducción implica. Esto hace muy probable que después de la gran inversión en tiempo y recursos en el desarrollo del DWH, se haga patente la necesidad de cambios fundamentales que traen consigo altos costos de desarrollo para la organización, poniendo en evidente peligro el éxito de todo el proyecto en sí y que podían ser evitados con una pronta detección en una temprana puesta en explotación de un primer avance del DWH.

Esta metodología para la construcción de un sistema de este tipo es frecuente a la hora de diseñar un sistema de información, utilizando las herramientas habituales como el esquema Entidad/Relación pero al tener un enfoque global, es más difícil de desarrollar en un proyecto sencillo, pues estamos intentando abordar el “todo”, a partir del cual luego iremos al “detalle”. Esta es otra de las restricciones que trabajan en contra de la metodología de Inmon ya que implica un consumo de tiempo mayor, teniendo como consecuencia que muchas empresas se inclinen por usar metodologías con las que obtengan resultados tangibles en un espacio menor de tiempo.

3.2. Metodologías más empleadas dentro del modelo bottom-up

3.2.1. Rapid Warehousing Methodology

Entre las metodologías que se emplean actualmente está la técnica "*Rapid Warehousing Methodology (RWM)*" propuesta por el Instituto SAS (EEUU) [4], en el año 1998. Dicha metodología es iterativa y está basada en el desarrollo incremental del proyecto de DWH dividido en cinco fases:

Figura 4: Metodología Rapid Warehousing

A continuación se describe con mayor detalle cada una de las fases de la metodología.

3.2.1.1. Definición de los objetivos

En esta fase se definirá el equipo de proyecto que debe estar compuesto por representantes del departamento informático y de los departamentos usuarios del DWH, además de la figura de jefe de proyecto.

Se definirá el alcance del sistema y cuáles son las funciones que el DWH realizará como suministrador de información de negocio estratégica para la empresa. Se definirán así mismo, los parámetros que permitan evaluar el éxito del proyecto.

3.2.1.2. Definición de los requerimientos de información

Durante esta fase se mantendrán sucesivas entrevistas entre los representantes del departamento usuario final y los representantes del departamento de informática. Se realizará el estudio de los sistemas de información existentes, que ayudarán a comprender las carencias actuales y futuras que deben ser resueltas en el diseño del DWH.

Asimismo, en esta fase el equipo de proyecto debe ser capaz de validar el proceso de entrevistas y reforzar la orientación de negocio del proyecto. Al finalizar esta fase se obtendrá el documento de definición de requerimientos en el que se reflejarán no solo las necesidades de información de los usuarios, sino cual será la estrategia y arquitectura de implantación del DWH.

3.2.1.3. Diseño y modelización

Los requerimientos de información identificados durante la anterior fase proporcionarán las bases para realizar el diseño y la modelización del DWH.

En esta fase se identificarán las fuentes de los datos (sistema operacional, fuentes externas,..) y las transformaciones necesarias para, a partir de dichas fuentes, obtener el modelo lógico de datos del DWH. Este modelo estará formado por entidades y relaciones que permitirán resolver las necesidades de negocio de la organización.

3.2.1.4. Implementación

La implantación de un DWH lleva implícitos los siguientes pasos:

- Extracción de los datos del sistema operacional y transformación de los mismos.

- Carga de los datos validados en el DWH. Esta carga deberá ser planificada con una periodicidad que se adaptará a las necesidades de refresco detectadas durante la fase de diseño del nuevo sistema.

- Explotación del DWH mediante diversas técnicas dependiendo del tipo de aplicación que se dé a los datos. Entre las técnicas más habituales podemos encontrar las siguientes:

    - Query & Reporting
    - On-line analytical processing (OLAP)
    - Executive Information System (EIS) ó Información de gestión
    - Decission Support Systems (DSS)
    - Visualización de la información

3.2.1.5. Revisión

La construcción del DWH no finaliza con la implantación del mismo, sino que es una tarea iterativa en la que se trata de incrementar su alcance aprendiendo de las experiencias anteriores.

Después de implantarse, debería realizarse una revisión del DWH planteando preguntas que permitan, después de los seis o nueve meses posteriores a su puesta en marcha, definir cuáles serían los aspectos a mejorar o potenciar en función de la utilización que se haga del nuevo sistema.

3.2.1.6. Gestión del Proyecto

La gestión del proyecto debe encargarse de la coordinación y ejecución de las distintas fases que conforman la construcción e implantación de un DWH. Este proceso se tiene que apoyar en una metodología específica para este tipo de trabajos, si bien es más importante que la elección de la mejor de las metodologías, el realizar un control para asegurar el seguimiento de la misma.

En las fases que se establezcan es fundamental incluir una fase de formación en la herramienta utilizada, para un máximo aprovechamiento de la aplicación. Seguir los pasos de la metodología y comenzar el DHW por un área específica de la empresa permitirá obtener resultados tangibles en un corto espacio de tiempo.

3.2.2. Metodología Kimball – Ciclo de Vida

Ralph Kimball es el autor considerado como el "Gurú" del DWH junto con Bill Inmon. Su metodología se ha convertido en el estándar de facto en el área de apoyo a las decisiones empresariales.

En el año 1998 dicha metodología se recoge como proceso a seguir en el desarrollo de un DWH con el libro: "*The Data Warehouse Lifecycle Toolkit*" [5]

La siguiente figura muestra de forma esquemática las fases que componen la metodología propuesta por Kimball y los siguientes apartados resumen el contenido de cada una de las fases.

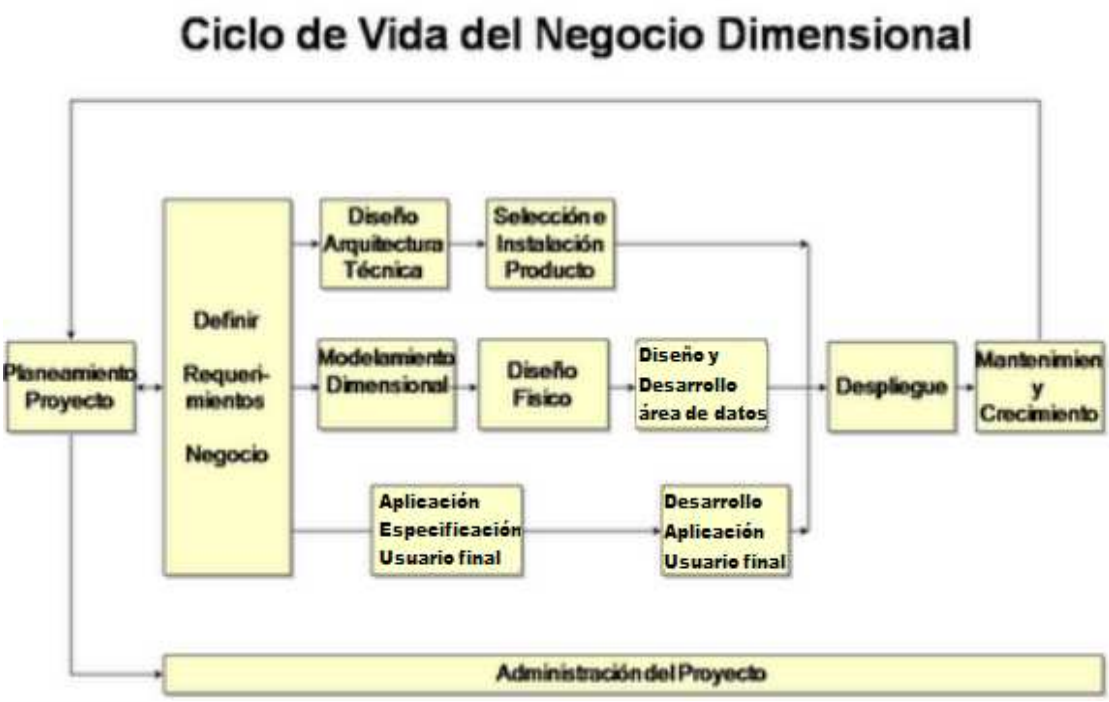

Figura 5: Ciclo de vida de la metodología de Ralph Kimball

3.2.2.1. Planificación del Proyecto

La planificación busca identificar la definición y el alcance del proyecto de DWH, incluyendo las justificaciones del negocio y las evaluaciones de factibilidad.

Esta etapa se concentra sobre la **definición** del proyecto. Según sentencia Kimball: *"Antes de comenzar un proyecto de data warehouse o data mart, hay que estar seguro si existe la demanda y de dónde proviene. Si no se tiene un usuario sólido, posponga el proyecto"*.

Como metodología, en esta etapa propone identificar el alcance preliminar basándose en los **requerimientos del negocio** y no en fechas límites, construyendo la justificación del proyecto en términos del negocio.

A nivel de planificación del proyecto se establece la identidad del mismo, el personal (los usuarios, gerentes del proyecto, equipo del proyecto), desarrollo del plan del proyecto, el seguimiento y la monitorización.

3.2.2.2. Definición de los Requerimientos del Negocio

Un factor determinante en el éxito de un proceso de DWH es la interpretación correcta de los diferentes niveles de requerimientos expresados por los distintos grupos de usuarios.

La técnica utilizada para revelar los requerimientos de los analistas del negocio difiere de los enfoques tradicionales guiados por los datos. Los diseñadores de los DWH deben entender los factores claves que guían el negocio para determinar efectivamente

los requerimientos y traducirlos en consideraciones de diseño apropiadas.

Los usuarios finales y sus **requerimientos** impactan **siempre** en la implementación de un DWH. Según la perspectiva de Kimball, los requerimientos del negocio se posicionan en el **centro** del *"universo del Data Warehouse"*. Como destaca siempre el autor, los **requerimientos del negocio deben determinar el alcance** del DWH (qué datos debe contener, cómo deben estar organizados, cada cuánto tiempo debe actualizarse, quiénes y desde dónde accederán, etc.).

3.2.2.3. Modelado Dimensional

La definición de los requerimientos del negocio determina los datos necesarios para cumplir los requerimientos analíticos de los usuarios. Diseñar los modelos de datos para soportar estos análisis requiere un enfoque diferente al usado en los sistemas operacionales. Básicamente, se comienza con una matriz donde se determina la dimensionalidad de cada indicador y luego se especifican los diferentes grados de detalle dentro de cada concepto del negocio, así como la granularidad de cada indicador y las diferentes jerarquías que dan forma al modelo dimensional del negocio (MDN) o mapa dimensional.

3.2.2.4. Diseño Físico

El diseño físico de la base de datos se focaliza sobre la selección de las estructuras necesarias para soportar el diseño lógico. Un elemento principal de este proceso es la definición de estándares del entorno de la base de datos. La indexación y las estrategias de particionamiento se determinan también en esta etapa.

En la estrategia de particionamiento o agregación, el DWH tiene, y debe tener, todo el detalle de información en su nivel atómico. Así, por poner algún ejemplo, en los sectores de telecomunicaciones o banca es habitual encontrarse con DWH con miles de millones de registros. Sin embargo, la mayoría de consultas no necesitan acceder a un nivel de detalle demasiado profundo. Un jefe de producto puede estar interesado en los totales de venta de sus productos mes a mes, mientras que el jefe de área consulta habitualmente la evolución de ventas de sus zonas. Incluso con el uso de índices, la compresión de las tablas, o con una inversión millonaria en hardware, estas consultas habituales deberían leer, agrupar y sumar decenas de millones de registros, lo que repercutiría directamente en el tiempo de respuesta y en el descontento de los usuarios. Por tanto, muchas veces lo más complicado será realizar la correcta elección de las tablas agregadas necesarias. De nada sirve crear muchos agregados si estos no se utilizan, por lo que es necesario conocer las consultas habituales de los usuarios para hacer la selección de las tablas agregadas.

La solución ante estas situaciones pasa siempre por la preparación de tablas agregadas. Estas tablas deben ser versiones reducidas de las dimensiones asociadas con la granularidad de la tabla de hechos y añaden los indicadores de las tablas de detalle a un nivel superior. Por ejemplo, las ventas podrían precalcularse a nivel mensual, o por cliente, o por producto. De esta manera, las consultas típicas del jefe de producto o del jefe de área podrían ejecutarse en pocos segundos, sin necesidad de acceder a la tabla de

ventas detalladas. La existencia de estas tablas agregadas debe ser completamente transparente para el usuario de negocio. Es decir, tanto el jefe de área como el jefe de producto trabajarán con el indicador "Ventas", y la herramienta de BI hará el resto.

Por otro lado, en la estrategia de indexación los índices son estructuras opcionales optimizadas y orientadas a conjuntos de operaciones. Según Ralph Kimball, las tablas de dimensión deben tener un único índice sobre las claves primarias y sería recomendable que el índice estuviera compuesto de un único atributo. Además recomienda el uso de índices de tipo árbol-B en atributos de alta cardinalidad y aplicar los índices de mapas de bits en atributos de cardinalidad media o baja.

La clave principal de la tabla de hechos es casi siempre un subconjunto de las claves externas, de manera que se elegirá un índice concatenado de las principales dimensiones de la tabla de hechos y dado que muchas consultas tienen relación con la dimensión fecha, ésta debería liderar el índice definido. Además, el atributo fecha en la primera posición permitirá aumentar la velocidad de los procesos de carga de datos que se agrupan por fecha y, dado que la mayoría de los optimizadores de consulta de los sistemas de gestión de bases de datos permiten que se utilice más de un índice a la hora de resolver una consulta, es posible construir diferentes índices en las demás claves ajenas de la tabla de hechos.

3.2.2.5. Diseño y Desarrollo de la Presentación de Datos

Esta etapa es típicamente la más subestimada de las tareas en un proyecto de DWH. Las principales actividades de esta fase del ciclo de vida son: *la extracción, la transformación y la carga* (*ETL process*). Se definen como procesos de extracción aquellos requeridos para obtener los datos que permitirán efectuar la carga del Modelo Físico diseñado. Así mismo, se definen como procesos de transformación los procesos para convertir o recodificar los datos fuente a fin de poder efectuar la carga efectiva del Modelo Físico. Por otra parte, los procesos de carga de datos son los procesos requeridos para poblar el DWH.

Todas estas tareas son altamente críticas pues tienen que ver con la materia prima del DWH: los datos. La desconfianza y pérdida de credibilidad del DWH provocará efectos inmediatos e inevitables si el usuario se encuentra con información inconsistente. Es por ello que la calidad de los datos es un factor determinante en el éxito de un proyecto de DWH. Es en esta etapa donde deben sanearse todos los inconvenientes relacionados con la calidad de los datos fuente. Para cumplir con estas premisas es necesario tener en cuenta ciertos parámetros a la hora de desarrollar las tablas de dimensión y la tabla de hechos.

3.2.2.6. Diseño de la Arquitectura Técnica

Los entornos de DWH requieren la integración de numerosas tecnologías. Se deben tener en cuenta tres factores: los requerimientos del negocio, los actuales entornos técnicos y las directrices técnicas y estratégicas futuras planificadas por la compañía para poder establecer el diseño de la arquitectura técnica del entorno de DWH.

Algunos equipos de trabajo no entienden las ventajas de una arquitectura y tienen la sensación de que las tareas son demasiado opacas, por lo que entienden su diseño como una distracción y un obstáculo para el progreso del DWH, así que optan por omitir el diseño de la arquitectura. Sin embargo, hay otros equipos de trabajo que dedican un tiempo demasiado grande para el diseño arquitectónico. El autor Ralph Kimball recomienda no irse a ninguno de los dos extremos para hacerlo de una manera intermedia. Para ello propone un proceso de 8 pasos para asegurar un correcto diseño arquitectónico sin extenderse demasiado en el tiempo.

- ***Establecer un Grupo de Trabajo de Arquitectura:***

Es muy útil disponer de un pequeño grupo de trabajo de dos a tres personas que se centren en el diseño de la arquitectura. Por lo general, es el arquitecto técnico, trabajando con los datos de diseño, el que estará al frente de este grupo de trabajo. Este grupo necesita establecer sus estatutos y la línea de prestaciones en el tiempo. También es necesario educar al resto del equipo sobre la importancia de una arquitectura.

- ***Requisitos relacionados con la arquitectura***

La arquitectura se crea para apoyar las necesidades del negocio, la intención no es comprar más productos. En consecuencia, el elemento fundamental para el proceso de diseño de la arquitectura proviene de los requerimientos de negocio obtenidos en esa fase de definición. El enfoque principal es descubrir las implicaciones arquitectónicas asociadas a las necesidades críticas del negocio, por lo que además de aprovechar la definición de los requisitos del proceso de negocio, también se llevan a cabo entrevistas adicionales dentro de la organización para comprender la normativa vigente dentro del marco tecnológico, instrucciones técnicas previstas y los límites no negociables.

- ***Documento de requisitos arquitectónicos***

Una vez definidos los requerimientos de negocio y llevado a cabo las entrevistas suplementarias es momento de documentar las conclusiones. La forma de hacerlo ha de ser sencilla pues el objetivo es tener una lista con cada requisito de negocio que tiene impacto en la arquitectura.

- ***Desarrollo de un modelo arquitectónico de alto nivel***

Una vez que los requisitos de la arquitectura se han documentado es hora de empezar a formular modelos para apoyar las necesidades identificadas. Para ello se dividen los equipos de trabajo según los componentes principales, como el acceso a datos, metadatos y la infraestructura. A partir de aquí, los equipos definen y refinan el modelo arquitectónico de alto nivel.

- ***Diseño y especificación de los subsistemas***

Una vez llegados a este punto es momento de hacer un diseño detallado de los subsistemas. Para cada componente, el grupo de trabajo diseña una lista con las capacidades necesarias de dicho componente. Por otro lado se tienen en cuenta las necesidades de seguridad, así como la infraestructura física y las necesidades de configuración. En algunos casos, las opciones de infraestructura, tales como el hardware del servidor y el software de base de datos, están predeterminados por la propia empresa. El tamaño, escalabilidad, rendimiento y flexibilidad son factores

clave a considerar al determinar el papel de los cubos OLAP en el conjunto de la arquitectura técnica.

- ***Determinar las fases de aplicación de la Arquitectura***

Es probable que no se puedan poner en práctica todos los aspectos de la arquitectura técnica a la vez. Algunos no son negociables, mientras que otros se pueden aplazar a una fecha posterior; éstos, son los requisitos de negocios para establecer las prioridades de la arquitectura.

- ***Documento de la Arquitectura Técnica***

Se debe de documentar la arquitectura técnica, incluyendo las fases de la implementación prevista. El documento de arquitectura incluirá información adecuada de manera que los profesionales cualificados puedan proceder con la construcción del sistema.

- ***Revisar y finalizar la Arquitectura Técnica***

El plan de la arquitectura se debe comunicar con diferentes niveles de detalle: equipo de proyecto, sponsor y director del proyecto. Tras la revisión, la documentación debe ser actualizada y utilizada inmediatamente en el proceso de selección del producto.

3.2.2.7. Selección de Productos e Instalación

Utilizando el diseño de arquitectura técnica como marco es necesario evaluar y seleccionar los componentes específicos de la arquitectura, como la plataforma de hardware, el motor de base de datos, la herramienta de ETL, las herramientas de acceso, etc.

Una vez evaluados y seleccionados los componentes determinados se procede con la instalación y prueba de los mismos en un ambiente integrado de DWH. Para ello es necesario tener en cuenta una serie de premisas que recomienda el autor de esta metodología:

- ***Comprender el proceso de compras corporativas.***

El primer paso antes de seleccionar nuevos productos es entender el hardware y el software interno, así como los procesos de aprobación de compras por parte de la empresa. Los gastos deben ser aprobados por el departamento correspondiente de la empresa.

- ***Elaborar una matriz de evaluación del producto.***

Con el plan de la arquitectura como punto de partida se desarrolla una matriz de evaluación empleando, por ejemplo, hojas de cálculo en donde se identificarán los criterios de evaluación, junto con factores de ponderación para indicar su importancia. Cuanto más específico sea el criterio, mejor. Estos criterios podrían incluir la funcionalidad, arquitectura técnica, características del software, impacto en las infraestructuras y viabilidad de los proveedores.

- ***Realizar investigación de mercados.***

Los compradores deben estar informados cuando van a seleccionar los productos. Esto significa realizar una amplia investigación del mercado para entender mejor a los vendedores y sus ofertas. La solicitud de información es una herramienta clásica de evaluación de productos.

- ***Filtrar opciones y realizar evaluaciones más detalladas.***

A pesar de la gran cantidad de productos disponibles en el mercado, sólo un pequeño número de los proveedores pueden satisfacer tanto nuestras necesidades técnicas como de funcionalidad. Mediante la comparación de resultados preliminares de la matriz de evaluación, debemos agrupar en una lista los proveedores sobre los que tomaremos la decisión. Con la lista de proveedores seleccionados se debe realizar un proceso de evaluación detallada, incluyendo si es posible otras instalaciones de tamaño similar sobre las que poder comparar a la hora de tomar una decisión.

- ***Manejo de un prototipo.***

Después de realizar la evaluación detallada, a veces hay un software ganador, a menudo basado en experiencias previas o relaciones con personal que provee el software. En muchas ocasiones, también puede surgir un producto debido a compromisos existentes con alguna de las empresas que ofertaban. En cualquier caso, cuando un candidato único aparece como la mejor opción, podemos evitar el uso de un prototipo con el consiguiente ahorro de tiempo y dinero. Si, por el contrario no existe una elección clara una vez que se llega a este momento, se debería llevar a cabo un prototipo con no más de dos productos, solicitando a los proveedores de software que proporcionen una solución con un pequeño conjunto de datos de muestra.

- ***Selección del producto, instalación y negociación.***

A la hora de seleccionar un producto en lugar de firmar inmediatamente con el proveedor, es necesario un periodo período de prueba en el que se ha de tener la oportunidad de utilizar el producto en su entorno real. A medida que la prueba llega a su fin se tiene la oportunidad de negociar una compra beneficiosa para todas las partes implicadas.

3.2.2.8. Especificación de Aplicaciones para Usuarios Finales

No todos los usuarios del DWH necesitan el mismo nivel de análisis. Es por ello que en esta etapa se identifican los roles o perfiles de usuarios para los diferentes tipos de aplicaciones necesarias en base al alcance de los perfiles detectados (gerencial, analista del negocio, vendedor, etc.)

3.2.2.9. Desarrollo de Aplicaciones para Usuarios Finales

A continuación de la especificación de las aplicaciones para usuarios finales, el desarrollo de las aplicaciones de los usuarios finales involucra configuraciones de los metadatos y construcción de reportes específicos.

Los usuarios acceden al DWH por medio de herramientas de productividad basadas en GUI (Graphical User Interface). De hecho existen multitud de estas herramientas con las que proveer a los usuarios. Las herramientas pueden incluir software de consultas, generadores de reportes, procesamiento analítico en línea o herramientas de Datamining dependiendo de los tipos de usuarios y sus requerimientos particulares. Sin embargo, una sola herramienta puede no satisfacer todos los requerimientos, por lo que quizás sea necesaria la integración de herramientas hechas bajo petición expresa de los usuarios para satisfacer sus necesidades de consulta sobre el DWH.

3.2.2.10. Implementación

La implementación representa la convergencia de la tecnología, los datos y las aplicaciones de usuarios finales accesibles para el usuario del negocio.

Hay varios factores extras que aseguran el correcto funcionamiento de todos estos elementos, entre ellos se encuentran la capacitación, el soporte técnico, la comunicación y las estrategias de feedback. Todas estas tareas deben tenerse en cuenta antes de que cualquier usuario pueda tener acceso al DWH.

3.2.2.11. Mantenimiento y crecimiento

Como se remarca siempre, la creación de un DWH es un proceso (de etapas bien definidas, con comienzo y fin, pero de naturaleza espiral) que acompaña a la evolución de la organización durante toda su historia. Se necesita continuar con las actualizaciones de forma constante para poder seguir la evolución de las metas por conseguir.

Al contrario de los sistemas tradicionales, los cambios en el desarrollo deben ser vistos como signos de éxito. Es importante establecer las prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa forma poder evolucionar y crecer.

Una vez que se ha construido e implantado el DWH no hay tiempo para el descanso, rápidamente debemos estar preparados para administrar el mantenimiento y crecimiento del mismo. Si bien las tareas pueden llegar a parecer similares a las tratadas en otras etapas del ciclo de vida, existe una diferencia clave: **los usuarios están ahora accediendo al DWH.**

3.2.2.12. Gestión del Proyecto

La gestión del proyecto asegura que las actividades del ciclo de vida se lleven a cabo de manera sincronizada. Como se indica en la figura 5, la gestión del proyecto acompaña todo el ciclo de vida. Entre sus actividades principales se encuentra la monitorización del estado del proyecto y el acoplamiento entre los requerimientos del negocio y las restricciones de los sistemas de información para poder manejar correctamente las expectativas en ambos sentidos.

4. Modelos y Arquitecturas habitualmente utilizados en un Data Warehouse

Un modelo es un conjunto de conceptos, reglas y convenciones que permiten describir y manipular los datos que queremos almacenar en una base de datos. En ellas se pueden usar distintos modelos de datos para describir la información con que operan. Un Modelo de Datos permite describir:

- La estructura de datos de la base: el tipo de los datos que hay en la base y la forma en que se relacionan.
- Las restricciones de integridad: Un conjunto de condiciones que deben cumplir los datos para reflejar correctamente la realidad deseada.

En entornos de DWH, los modelos más extendidos para representar los datos se basan en el hecho de que no todas las entidades tienen igual número de ocurrencias ni presentan igual curva de crecimiento en el volumen de dichos datos. Los DWH están más orientados a la consulta para apoyar la toma de decisiones que los sistemas operacionales que están más dedicados a trabajar con la operativa diaria de la empresa. Estas diferentes visiones del problema tienen una gran importancia en la manera de modelar.

4.1. Modelo en Estrella

El esquema en estrella es el más sencillo de los esquemas de almacenamiento de datos. Se llama así porque el diagrama se asemeja a una estrella, con los puntos que irradian desde un centro. El centro de la estrella consta de una o más tablas de hechos y los puntos de la estrella son las tablas de dimensiones, como se muestra en la figura 6 En concreto este esquema en estrella es ideal por su simplicidad y velocidad para ser usado en análisis multidimensionales como los DM, ya que permite acceder tanto a datos agregados como de detalle. Además, ofrece la posibilidad de implementar la funcionalidad de una base de datos multidimensional utilizando una clásica base de datos relacional.

El esquema en estrella consiste en estructurar la información en procesos, vistas y métricas a modo de estrella. En la tabla de hechos encontramos los atributos destinados al hecho que constituye el proceso de negocio a medir, es decir, sus métricas. Mientras, en las tablas de dimensión, los atributos se destinan a elementos de nivel (que representan los distintos niveles de las jerarquías de dimensión) y a atributos de dimensión (encargados de la descripción de estos elementos de nivel). En el esquema en estrella la tabla de hechos es la única tabla que tiene múltiples *joins* que la conectan con otras tablas. El resto de tablas del esquema (tablas de dimensión) únicamente hacen *join* con esta tabla de hechos. Las tablas de dimensión se encuentran además totalmente desnormalizadas, es decir, toda la información referente a una dimensión se almacena en la misma tabla.

La figura 6 muestra un ejemplo de diagrama utilizando el Modelo en Estrella obtenido de *"Data Warehousing Guide"* [6]

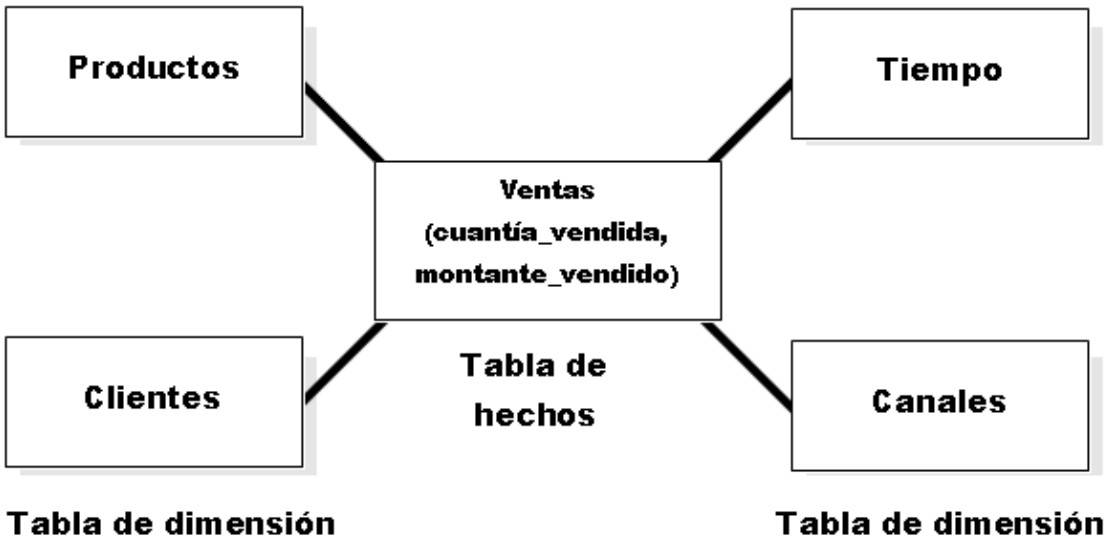

Figura 6: Ejemplo de un esquema en estrella

4.2. Modelo en copo de nieve (Snowflake)

El esquema en copo de nieve (*snowflake*) es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Por esta razón, la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas, y aparecen nuevas *join* o uniones entre tablas gracias a que las dimensiones de análisis se representan ahora en tablas de dimensión normalizadas. En la estructura dimensional normalizada, la tabla que representa el nivel base de la dimensión es la que hace *join* directamente con la tabla de hechos. La diferencia entre ambos esquemas (estrella y copo de nieve) reside entonces en la estructura de las tablas de dimensión. Para conseguir un esquema en copo de nieve se ha de tomar un esquema en estrella y conservar la tabla de hechos, centrándose únicamente en el modelado de las tablas de dimensión, que si bien en el esquema en estrella se encontraban totalmente desnormalizadas, ahora se dividen en subtablas tras un proceso de normalización.

Es posible distinguir dos tipos de esquemas en copo de nieve, un *snowflake* completo (en el que todas las tablas de dimensión en el esquema en estrella aparecen normalizadas) o un snowflake parcial (sólo se lleva a cabo la normalización de algunas de ellas). En la siguiente figura se observa el *modelo en copo de nieve* descrito obtenido de "*Data Warehousing Guide*" [6]:

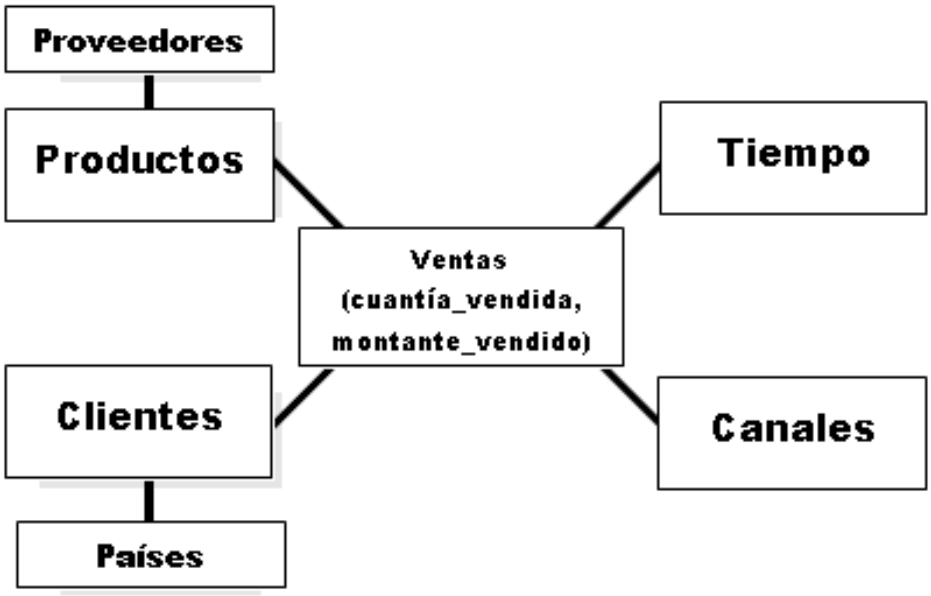

Figura 7: Ejemplo de un esquema en copo de nieve

4.3. Modelo dimensional

Aunque en los entornos de BI los esquemas de estrella son los más empleados, también se puede utilizar la tercera forma normal para su utilización en sistemas de almacenamiento datos.

El modelado de la Tercera Forma Normal (3FN) es una técnica clásica de las bases de datos relacionales que minimizan la redundancia de datos a través de la normalización de los datos. Cuando se compara con un esquema en estrella, un esquema 3FN tiene normalmente un mayor número de tablas debido al proceso de normalización.

Los esquemas 3FN se utilizan en los almacenes de datos grandes, especialmente en entornos con importantes requisitos de carga de datos que se utilizan para alimentar DM y ejecutar consultas de larga ejecución.

Las principales ventajas de los esquemas 3FN son las siguientes:

- Proporcionan un diseño de esquema neutral e independiente de cualquier aplicación.
- Puede requerir menos transformación de datos que otros esquemas como los esquemas en estrella.

En la figura 8 se puede ver un diseño de un esquema 3FN obtenido de "*Data Warehousing Guide*" [6]

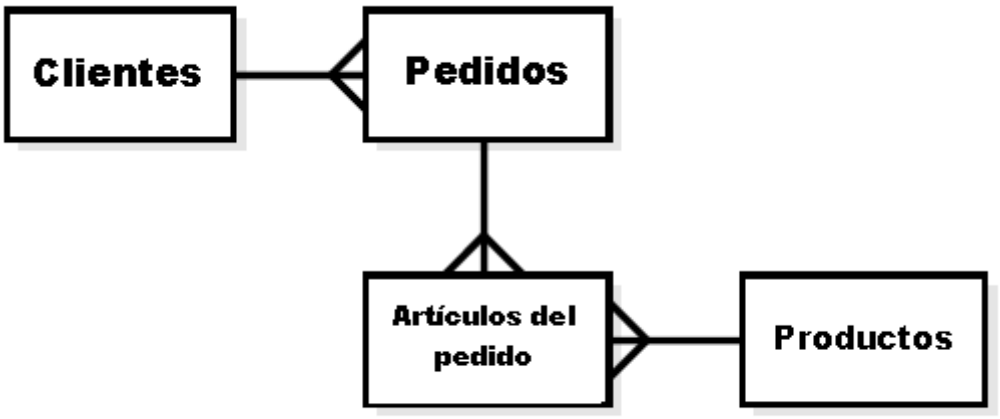

Figura 8: Ejemplo de un esquema en 3FN

4.4. Elección de un modelo

En el desarrollo de un DWH, la forma más natural de construir el diseño es mediante el modelo en estrella, donde sólo se establece una relación entre la tabla de hechos y las tablas de dimensiones. De esta manera se optimiza el rendimiento al mantener las consultas lo más simples que sea posible y proporcionar servicios rápidos con un tiempo de respuesta lo más corto posible, al almacenar toda la información acerca de cada nivel en pocas tablas.

Por otro lado, el esquema de copo de nieve proporciona gran simplicidad desde el punto de vista del usuario final. El argumento a su favor es que al estar normalizadas con las tablas de dimensiones se evita la redundancia de datos y con ello se ahorra espacio. Pero si tenemos en cuenta que hoy en día, el espacio en disco no suele ser un problema, y sí el rendimiento, se presenta como una mala opción en un DWH, ya que el hecho de disponer de más de una tabla por cada dimensión implica tener que realizar código más complejo para realizar una consulta que a su vez se ejecutará en un tiempo mayor, debido en parte al mayor número de uniones (*joins*) que habrá que realizar. Por tanto, se puede usar un esquema de copo de nieve en un DWH, aunque estos sean realmente grandes y complejos, pero nunca en sistemas donde el tiempo de respuesta sea un factor crítico para los usuarios. Similares problemas se pueden encontrar si se utiliza un diseño basado en el modelo relacional puro.

En resumen, si la principal preocupación del proyecto a realizar está marcada por las premisas de espacio y rendimiento, el esquema en estrella se presenta como la opción más aconsejable, pues permite indexar las dimensiones de forma individualizada sin que repercuta en el rendimiento de la base de datos en su conjunto.

5. Arquitectura de un Data Warehouse

Una de las razones por las que el desarrollo de un DWH crece rápidamente se debe a que realmente es una tecnología fácilmente entendible. De hecho, un DWH puede representar mejor que otros sistemas la compleja estructura de una empresa a la hora de administrar los datos gerenciales dentro de la organización. A fin de comprender cómo se relacionan todos los componentes involucrados en una estrategia de DWH es esencial tener una arquitectura de DWH.

5.1. ¿Qué posibles arquitecturas contempla un Data Warehouse?

La construcción del DWH se establece como un elemento crítico en el proceso de implantación de una herramienta de BI y por lo tanto resulta interesante revisar ciertos aspectos antes de abordar el diseño [7]:

- **Base de datos operacional, nivel de base de datos externos*:*** Los sistemas operacionales procesan datos para apoyar las necesidades operacionales críticas. Estos sistemas proporcionan una estructura de procesamiento eficiente para un gran número de transacciones comerciales bien definidas.

  Por otro lado, la meta del DWH es utilizar la información que se almacena en bases de datos operacionales y combinarla con la información que proviene de otras fuentes de datos, generalmente externas.

- **Nivel de acceso a la información*:*** El nivel de acceso a la información de la arquitectura DWH es el nivel del que el usuario final se encarga directamente. En particular, representa las herramientas que el usuario final normalmente usa día a día.

  Este nivel incluye el hardware y software involucrados en mostrar información en pantalla y emitir informes, hojas de cálculo, gráficos y diagramas para el análisis y presentación.

- **Nivel de acceso a los datos:** El nivel de acceso a los datos de la arquitectura DWH está relacionado con el nivel de acceso a la información en el nivel operacional.

  El nivel de acceso a los datos conecta DBMS's diferentes y sistemas de archivos sobre el mismo hardware y protocolos de red. Una de las claves de una estrategia DWH es proveer a los usuarios finales de un "acceso universal a los datos".

  El acceso a los datos universales significa que, teóricamente por lo menos, los usuarios finales sin tener en cuenta la herramienta de acceso a la información o la ubicación, deberían ser capaces de acceder a cualquier o todos los datos en la empresa que son necesarios para ellos, para hacer su trabajo. El nivel de acceso a los datos entonces es responsable de las interfaces entre las herramientas de acceso a la información y las bases de datos operacionales. En algunos casos,

esto es todo lo que un usuario final necesita. Sin embargo, en general, las organizaciones desarrollan un plan mucho más sofisticado para el soporte del DWH.

- **Nivel de directorio de datos (metadatos):** A fin de proveer el acceso a los datos universales, es absolutamente necesario mantener alguna forma de directorio de datos o metadatos. Los metadatos son información sobre los datos dentro de la empresa.

  Con el objeto de tener una base de datos totalmente funcional, es necesario tener una variedad de metadatos disponibles, información sobre las vistas de datos de los usuarios finales e información sobre las bases de datos operacionales. Idealmente, los usuarios finales deberían de acceder a los datos desde el DWH (o desde las bases de datos operacionales), sin tener que conocer dónde residen los datos o la forma en que están almacenados.

- **Nivel de gestión de procesos:** El nivel de gestión de procesos tiene que ver con la programación de diversas tareas que deben realizarse para construir y mantener el DWH y la información del directorio de datos. Este nivel depende del nivel de trabajo de los procesos encargados de mantener el DWH actualizado.

- **Nivel de mensaje de la aplicación:** El nivel de mensaje de la aplicación tiene que ver con el transporte de información alrededor de la red de la empresa. Puede usarse para aislar aplicaciones operacionales o estratégicas a partir de un formato de datos exacto, recolectar transacciones o los mensajes y entregarlos a una ubicación segura en un tiempo específico.

- **Nivel Data Warehouse (físico):** Es el repositorio central altamente flexible de información donde residen copias de los datos operacionales usados principalmente para fines estratégicos. En un DWH físico las copias de datos operacionales y/o externos se almacenan de forma que sea fácil acceder. Actualmente, los DWH se almacenan en plataformas cliente/servidor, pero también existen configuraciones sobre mainframes y/o equipos externos optimizados para su acceso mediante consulta.

- **Nivel de organización de datos:** El componente final de la arquitectura DWH es la organización de los datos. Incluye todos los procesos necesarios para seleccionar, editar, resumir, combinar y cargar datos en el DWH y para acceder a la información desde bases de datos operacionales y/o externas.

  La organización de datos involucra en muchas ocasiones una programación compleja, pero cada vez con mayor frecuencia se están creando herramientas de DWH para ayudar en este proceso.

En la figura 9 obtenida del Documento de Sistema Gerencial de Apoyo Universitario [9], se pueden observar las tres principales capas de la arquitectura de un DWH, es decir, el nivel de organización, el nivel de directorio y por último el nivel de gestión de procesos desde abajo hacia arriba en la imagen.

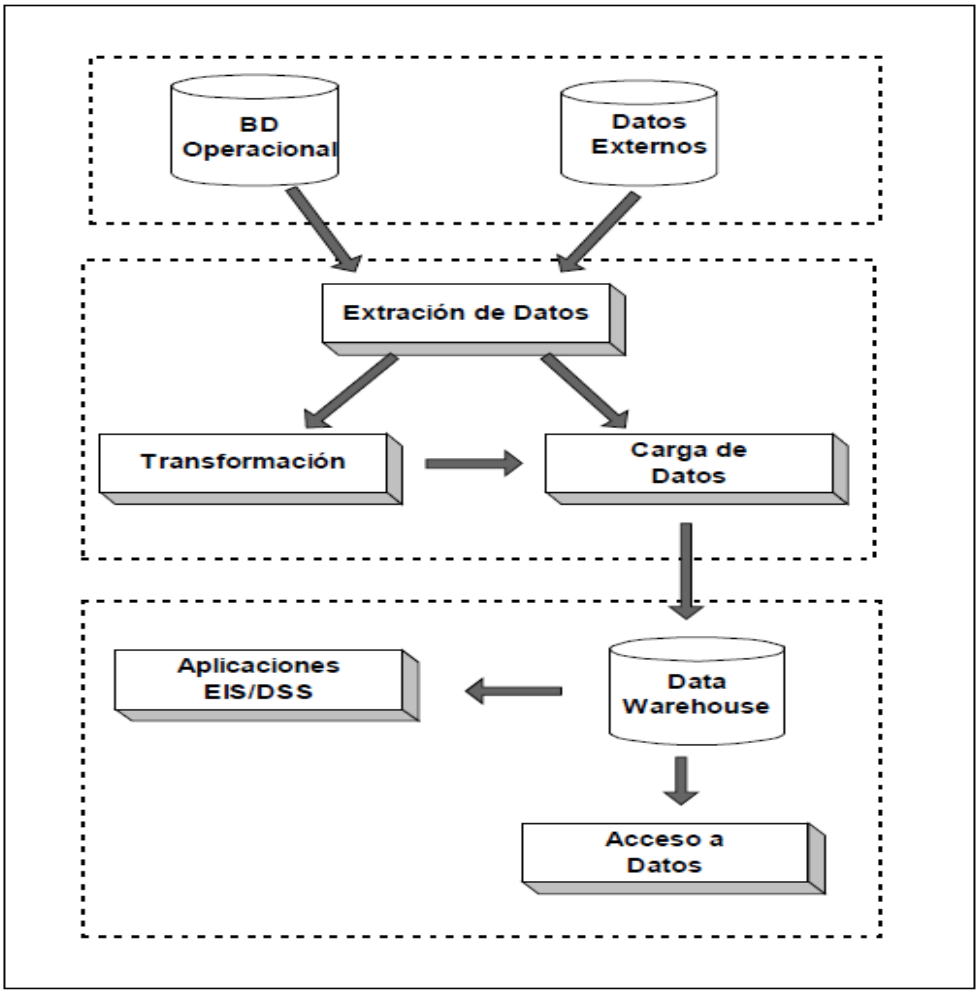

Figura 9: Estructura básica de un Data Warehouse

La forma en la cual se estructura el almacenamiento de datos en el DWH genera una clasificación respecto a la forma de implementar una arquitectura de DWH. La estructura adoptada para el DWH se debe realizar de la manera que mejor satisfaga las necesidades empresariales, siendo entonces dicha elección un factor clave en la efectividad del DWH. Así pues, una vez establecidos los conceptos básicos sobre los diferentes niveles existentes que conforman los DWH, nos encontramos con varios modelos de arquitectura, apareciendo tres como los más habituales según el manual "*Data Warehousing Guide*" [6]:

1. ■ Arquitectura Data Warehouse básica.
2. ■ Arquitectura Data Warehouse con área de organización.
3. ■ Arquitectura Data Warehouse con área de organización y Data Marts.

5.1.1. Arquitectura Data Warehouse Básica

Los usuarios finales acceden directamente a los datos desde diferentes sistemas de recursos a través del DWH como se observa en la figura 10 obtenida [6].

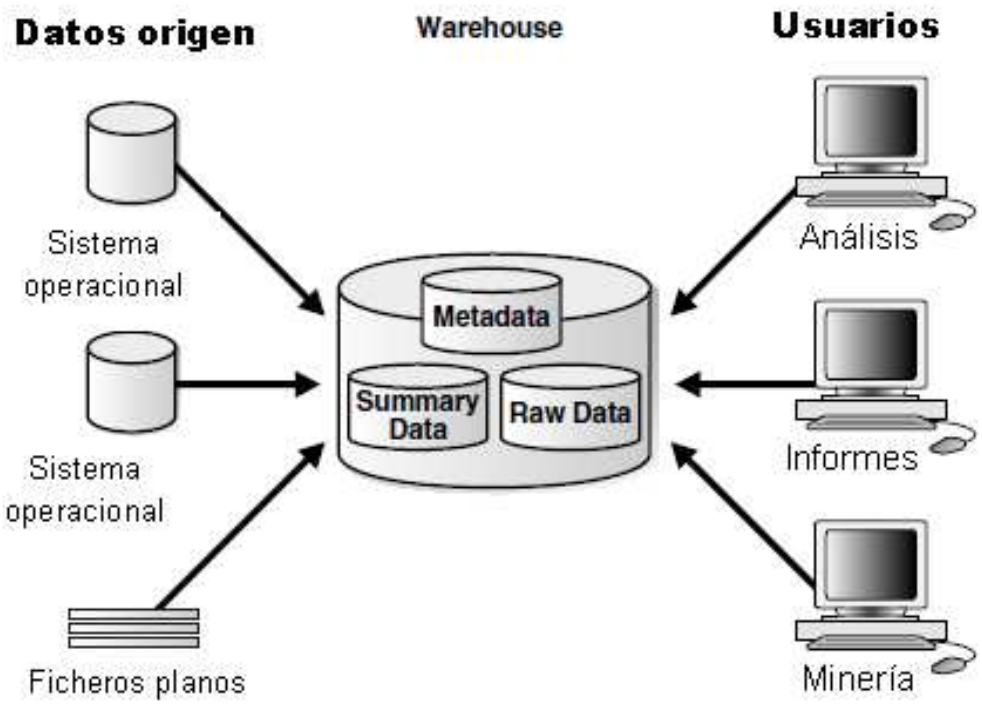

Figura 10: Arquitectura Data Warehouse básica

En la figura se puede observar cómo están presentes los metadatos y los datos sin procesar *(raw data)* propios de un entorno operacional. Por otro lado, apreciamos la aparición de los *summary data* ó datos resumen, datos muy importantes en los entornos de DWH porque permiten calcular con antelación las operaciones más costosas.

En los sistemas de soporte de decisiones (DSS) se hace un mayor uso de los datos resumen ó *summary data* ya que son mucho menos voluminosos y mucho más fáciles de gestionar que los datos detallados. Desde la perspectiva del acceso y presentación, el resumen de datos es ideal para una buena gestión ya que representa una base sobre la que construir futuros análisis y los ya existentes no se repitan. Por estos motivos, el resumen de datos forma parte integral de los entornos DSS.

5.1.2. Arquitectura Data Warehouse con Área de Organización

Este modelo arquitectónico se basa en el hecho de que para que se introduzcan los datos operativos en un DWH, es necesario que antes se limpien y procesen para su posterior almacenamiento. Es posible realizar estas acciones mediante programación, pero los DWH incorporan un área de organización en el que simplificar los resúmenes generales y la gestión de los almacenes como se observa en la figura 11 obtenida de [6].

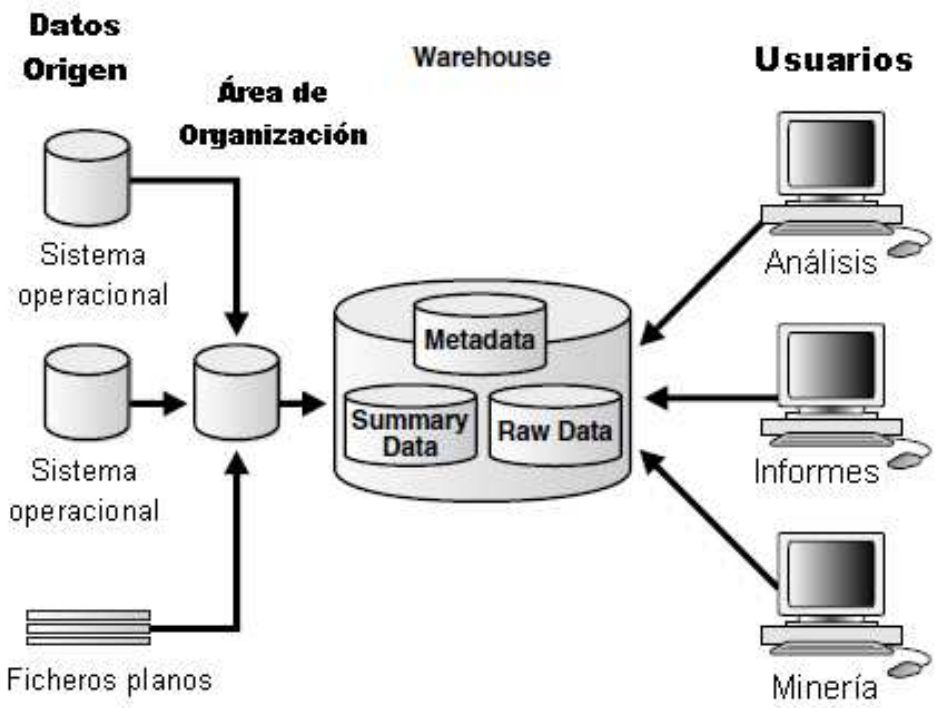

Figura 11: Arquitectura de un Data Warehouse con área de organización

El área de organización es donde se colocan los datos en tránsito, por lo general procedentes de la capa de procesamiento ETL previa. En su forma más simple, son datos que han sido resumidos, de manera que la información detallada se ha sintetizado hasta el menor detalle. La forma en que estos datos se almacenen físicamente depende de la preferencia de los analistas y administradores de bases de datos, aunque en muchos casos se modelan siguiendo un esquema de estrella.

Para obtener estos *summary data* puede que se hayan utilizado herramientas de transformación complejas (ETL), obteniendo el nivel de detalle adecuado a las necesidades del usuario y de fácil acceso a los datos.

5.1.3. Arquitectura Data Warehouse con Área de Organización y Data Marts

Este modelo combina ideas de los dos anteriores, de forma que se implementa tanto el almacén empresarial como los departamentales. Se dispone de una base de datos, generalmente de detalle o de información común a todos los usuarios y además cada departamento, área o línea de negocio dispone de su propia base de datos. Es un modelo arquitectónico bastante común y permite personalizar la arquitectura del DWH según los diferentes grupos que existan en la organización. Esta arquitectura se realiza mediante la incorporación de DM al sistema global debido a que son sistemas diseñados para una determinada línea de negocio como se observa en la figura 12 obtenida de [6].

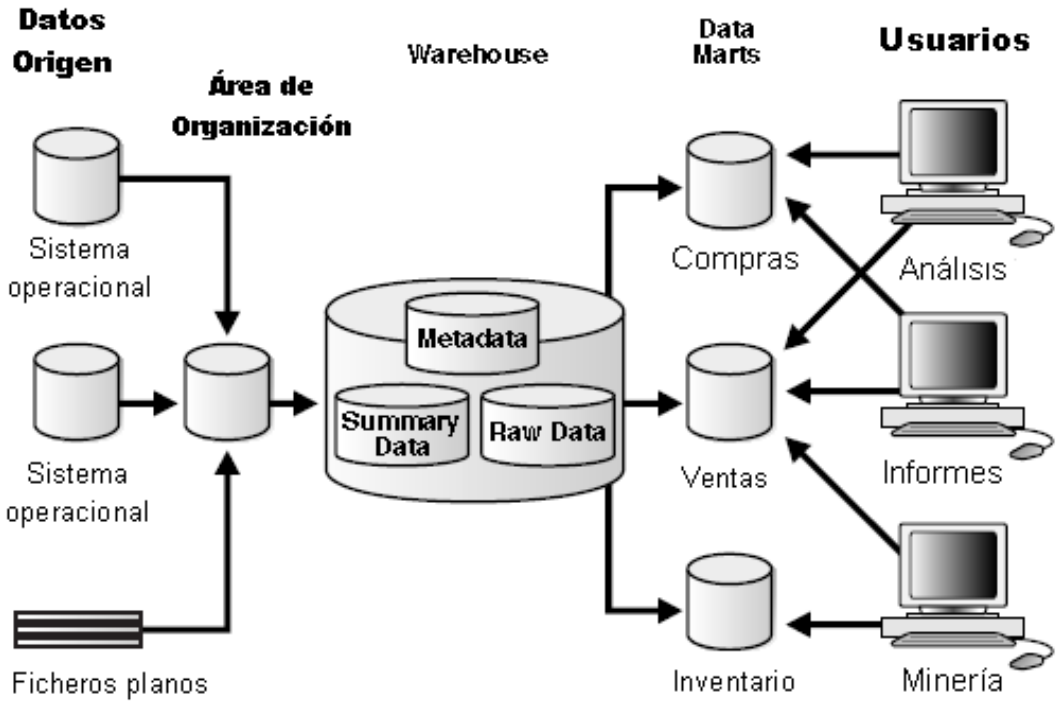

Figura 12: Arquitectura de un Data Warehouse con área de organización y Data Marts

Los DM se crean con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades. Esto se debe a que son sistemas orientados a consulta, en los que se ejecutan procesos batch de carga de datos con una frecuencia baja y conocida.

Como se puede observar en la figura 12, los DM son consultados por los diferentes usuarios y, mediante herramientas OLAP, ofrecen una visión multidimensional de la información.

Parte II: Data Mart para el Seguimiento Académico de Alumnos en el Entorno Universitario

1. Definición de los Requerimientos del Negocio

El interés de este proyecto se va a centrar en analizar las etapas por las que puede pasar el alumnado de las diferentes carreras universitarias que se imparten en cada uno de los campus universitarios: desde su comienzo en los estudios hasta llegar a terminar la carrera, pasando por la superación de cada uno de los cursos. También será objeto de análisis las razones por las que el alumno tiene dificultades a la hora de superar un curso en concreto de la carrera pero sin embargo otros los supera sin mucha dilación. Otra situación que se produce con ciertos alumnos es que ni abandonan ni terminan la carrera pero su expediente debe mantenerse.

Un campus universitario se compone de un gran número de carreras en las que podemos realizar el estudio y además, las distintas carreras se encuadran en diferentes áreas en las que se desarrollan: ciencias sociales, ciencias experimentales, ingenierías, ciencias jurídicas, etc. Esto permite que el alumno, en caso de abandonar su carrera universitaria, pueda matricularse dentro del mismo campus o en otro diferente en otras áreas, pudiendo ser también objeto del estudio en cuestión. Con estos datos, se podrán sacar patrones en la toma de decisiones del alumnado dentro de cada área y los porcentajes de éxito, abandono, cambio de titulación o cambio de área.

Por otro lado, una carrera incluye un gran número de asignaturas a lo largo de la misma. Y dentro de éstas, aunque la titulación esté orientada a un área concreta, por ejemplo ingeniería informática, sus asignaturas pueden tratar áreas que complementan la formación de dicho alumno como las orientadas a la dirección de empresas.

El hecho de que se impartan diferentes tipos de asignaturas dentro de una carrera puede hacer que los alumnos tengan mayor o menor número de dificultades para superarlas, cuestión que será interesante analizar. En relación con este asunto debemos incluir el apartado de las asignaturas optativas de la propia carrera, ya que basándose en la información de años anteriores, los alumnos pueden ir escogiendo las más sencillas o complicadas dependiendo del tipo de alumno que se analice.

En cuanto al tipo de alumno, está claro que no todos tienen las ideas claras sobre la carrera en que se matriculan, lo que hace que en un cierto número de casos, la tasa de abandono del alumno en ciertas carreras universitarias debido a esta razón no sea baja.

En el caso de los alumnos que terminan sus estudios es también interesante saber la tasa de aprobados y las calificaciones dependiendo de la carrera, el tipo de alumno que se quiera observar (según su género o según la edad), el número de matriculados en la asignatura, el cuatrimestre en el que se imparten las asignaturas y la convocatoria del examen. Estos últimos elementos en cuanto a la influencia que pueden tener las asignaturas en el estudio del seguimiento. Todo ello puede influir en las tasas de aprobados, ya que puede haber alumnos que saquen más rendimiento al estudio para una u otra convocatoria.

Como conclusión, a partir de las hipótesis anteriores el proyecto debería ofrecer información útil sobre el ciclo académico de los alumnos, considerando como tal el paso por una serie de etapas, desde la matrícula inicial hasta la finalización, abandono o bloqueo (situación en la que el alumno ni termina ni abandona). La información sobre el ciclo académico permitirá detectar posibles cuellos de botella, porcentajes de éxito y abandono, asignaturas y cursos más difíciles en cuanto al número de alumnos que tardan más de un año en completar esas asignaturas y cursos y cualquier otra cuestión que permita mejorar el grado de satisfacción de los alumnos y la calidad de los estudios.

2. ¿Qué metodología emplear para construir un DWH en el entorno de alumnos universitarios?

Una vez descritas las metodologías más habituales que se emplean para desarrollar un DWH, podemos seleccionar cuál de ellas proporciona mayores ventajas frente a las demás en el estudio académico de alumnos universitarios.

Por documentación y casos implantados en un gran número de sectores diferentes (banca, ventas, comunicaciones, educación, etc) en los que poder apoyar la toma de decisiones, parece más asequible una elección entre las dos grandes metodologías conocidas: la de Ralph Kimball y la de Bill Inmon, por lo que en los apartados siguientes trataremos de comprobar cual se adapta mejor a las necesidades del proyecto. Además, estas dos metodologías sirven como base para otras propuestas que utilizan gran parte de los conceptos descritos en ellas.

2.1. ¿Porqué usar la metodología de Kimball?

La metodología de Kimball conduce a una solución completa en una cantidad de tiempo relativamente pequeña. Además, debido a la gran cantidad de documentación que se puede encontrar y a los numerosos ejemplos aportados en diferentes entornos, permite encontrar una respuesta a casi todas las preguntas que puedan surgir, sobre todo cuando no se dispone de la experiencia previa necesaria.

Por otro lado, este tipo de metodología **bottom-up** permite que, partiendo de cero, podamos empezar a obtener información útil en cuestión de días y después de los prototipos iniciales, comenzar el ciclo de vida normal que nos ofrezca una solución completa de BI.

Los DM resultantes son fácilmente consultables tanto para los desarrolladores como para los usuarios finales. La relación directa entre los hechos y dimensiones conceden a cualquier usuario la posibilidad de construir consultas muy sencillas, la mayoría de las veces sin tener a mano la documentación de los metadatos.

La metodología de Kimball es ideal para los primeros pasos de la implantación de BI a un cliente, cuando la complejidad del almacenamiento de datos no es demasiado grande y donde la infraestructura del BI se encarga de los datos procedentes de un número limitado de fuentes. Sin embargo, cuando el almacén de datos adquiere complejidad, entonces es peligroso forzar el desarrollo de esta metodología. En el mundo del BI, cuando las cosas adquieren gran complejidad, es el momento de introducir nuevos enfoques al problema, como el propuesto por Inmon.

2.2. ¿Porqué usar la metodología de Inmon?

Al plantear esta metodología en el proceso de creación de un DWH es importante tener en cuenta la diferencia de visiones entre los dos autores citados. Para Bill Inmon el DWH es una parte de un sistema de BI dentro de una empresa, que tiene

un DWH y los DM obtienen su información a partir de este DWH. Por otro lado Ralph Kimball plantea que el DWH es un conglomerado de todos los DM dentro de una empresa y la información siempre se almacena de acuerdo al Modelo Dimensional.

Esto implica que, para Bill Inmon el desarrollo del DWH debe ser completo para su correcto funcionamiento mientras que la visión de Ralph Kimball nos permite desarrollar DM particulares que contengan la lógica de negocio en la que se está interesado en profundizar, permitiendo de esta manera que no se tenga que realizar, por ejemplo en el caso de estudio de este proyecto, el desarrollo de todas las problemáticas de negocio que se puedan contemplar en el proceso educativo universitario.

Por último, antes de decidir si utilizar o no la metodología de Inmon, es interesante plantear una serie de cuestiones para conocer la conveniencia de su uso ya que a priori sería más costosa de realizar que la metodología de Kimball.

Suponiendo que hemos partido de la metodología de Kimball y vamos viendo que nuestro desarrollo se complica a cada paso que damos, deberíamos plantearnos las siguientes preguntas:

- *¿Hay algún software que consulta nuestro almacén de datos de forma predecible?*
- *¿Podemos obtener opiniones generales sobre los datos a través de los DM existentes?*
- *¿Todos los datos están contenidos en los DM?*
- *¿Cuántos cambios habría que aplicar a los DM existentes si aparecen cambios en la estructura de los datos procedentes de las fuentes OLTP?*
- *¿Existe una necesidad de contar con detalle y profundidad algo que no podamos representar fácilmente con tablas de hechos y dimensiones?*

Tenemos que responder a estas preguntas cada vez que se quiera añadir una nueva característica al sistema. Cuando consideramos que la complejidad de toda la solución de BI se está complicando demasiado, poco a poco podemos vislumbrar que la estructura de Inmon es la solución para dar lugar a un almacén de datos más complejo.

En cualquier caso, si utilizamos la metodología de Inmon, la estructura final está siempre separada de los DM. Por lo tanto, el almacén de datos es un paso intermedio en el que se puede consolidar y limpiar los datos con el fin de alimentar a los DM.

2.3 Elección de la metodología de Kimball para el DWH en el entorno académico de alumnos universitarios

Teniendo en cuenta la información de los apartados anteriores y el interés de este proyecto en centrar el estudio en ciertos aspectos dentro del entorno universitario, la metodología de Ralph Kimball se ajusta más a lo que se quiere desarrollar al permitir la creación del DWH partiendo de los DM. En particular, debido a que el interés de este proyecto se enfoca en los aspectos académicos de los alumnos, se puede considerar que este planteamiento es útil desde el punto de vista de los departamentos universitarios que tratan con alumnos y sus progresos académicos, considerándose este grupo un subconjunto del total de departamentos y organismos de la universidad. De esta forma, el proyecto puede considerarse un primer acercamiento al BI académico a partir del estudio de un proceso de negocio concreto. La metodología de Kimball nos permite, por tanto, ofrecer soluciones en un plazo inferior al que resultaría de abordar un proyecto global destinado a toda la institución universitaria.

Por otro lado, la metodología de Kimball ofrece una clara exposición de las fases y actividades propias de cada fase, así como un buen número de ejemplos documentados en los cuales apoyarse cuando no se dispone de gran experiencia en el desarrollo de DM y DWH. Especialmente importante a este respecto son sus recomendaciones sobre el modelado dimensional.

3. Planificación del Proyecto

Como en todo proyecto software, es necesario realizar una planificación del mismo para tratar de garantizar que el desarrollo se realiza según las fechas previstas. A continuación se presenta la planificación prevista para el proyecto de DM que se está realizando:

	Nombre de tarea	Comienzo real	Fin real	% completado	Duración real	Trabajo real
1	Definición de los requerimientos de negocio	lun 23/11/09	vie 27/11/09	100%	10 días	40 horas
2	**Modelado dimensional**	**lun 07/12/09**	**lun 08/03/10**	**100%**	**132 días**	**520 horas**
3	Definir modelo de negocio	lun 07/12/09	vie 18/12/09	100%	20 días	80 horas
4	Definir el grano	mar 22/12/09	lun 28/12/09	100%	10 días	40 horas
5	Elegir las dimensiones	mar 29/12/09	lun 18/01/10	100%	30 días	120 horas
6	Identificar los hechos	mar 19/01/10	lun 01/02/10	100%	20 días	80 horas
7	Detallar las dimensiones	mar 02/02/10	lun 08/03/10	100%	50 días	200 horas
8	Diseño Físico del Data Warehouse	mar 09/03/10	mar 06/04/10	100%	42 días	168 horas
9	Diseño y Desarrollo de la presentación de datos	mié 07/04/10	mar 13/04/10	100%	10 días	40 horas
10	Diseño de la arquitectura técnica	mié 14/04/10	mar 04/05/10	100%	30 días	120 horas
11	Selección de productos e instalación	mié 05/05/10	mar 11/05/10	100%	10 días	40 horas
12	Especificación de aplicaciones para usuarios finales	mié 12/05/10	mar 18/05/10	100%	10 días	40 horas
13	Desarrollo de Aplicaciones para Usuarios Finales	mié 19/05/10	mié 26/05/10	100%	12 días	48 horas
14	Implementación	jue 27/05/10	mié 02/06/10	100%	10 días	40 horas
15	Mantenimiento y crecimiento	jue 03/06/10	mar 08/06/10	100%	8 días	32 horas
16	Gestión del proyecto	lun 23/11/09	mar 08/06/10	100%	284 días	0 horas

Figura 13: Planificación del proyecto

Como se puede observar en la figura 13, el mayor tiempo del proyecto se destina al modelado dimensional, fase que agrupa cuatro subtareas que son en las que más hincapié se hará en el presente documento. En la siguiente figura se puede observar las relaciones entre las diferentes tareas establecidas para la realización del proyecto, así como las dependencias entre unas y otras.

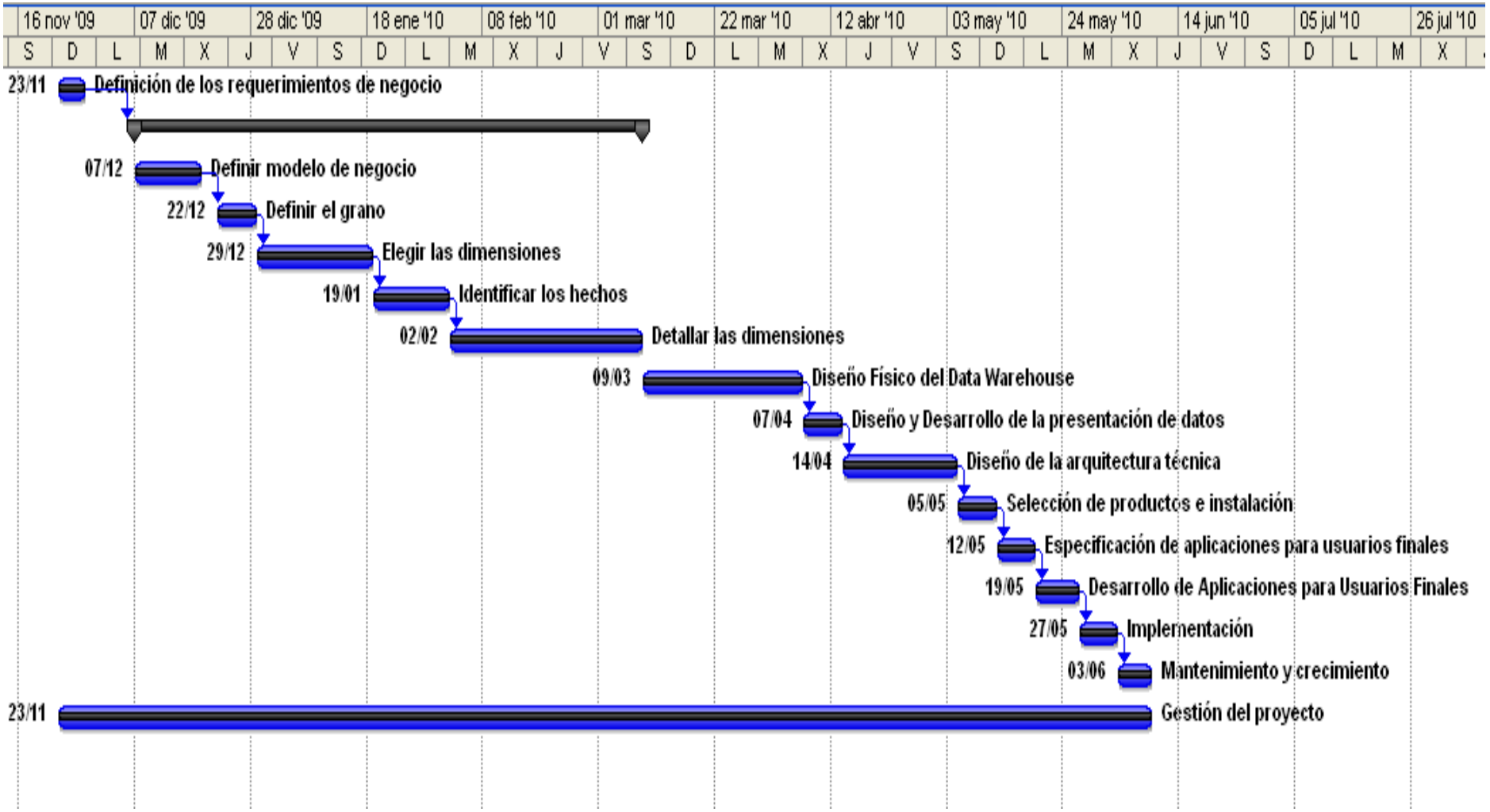

Figura 14: Planificación del proyecto de DM académico.

En los siguientes apartados se definen con detalle cada una de las fases del proyecto. Se incluye de nuevo la figura del ciclo de vida de la metodología de Kimball por cuestiones de claridad y facilidad de uso del documento.

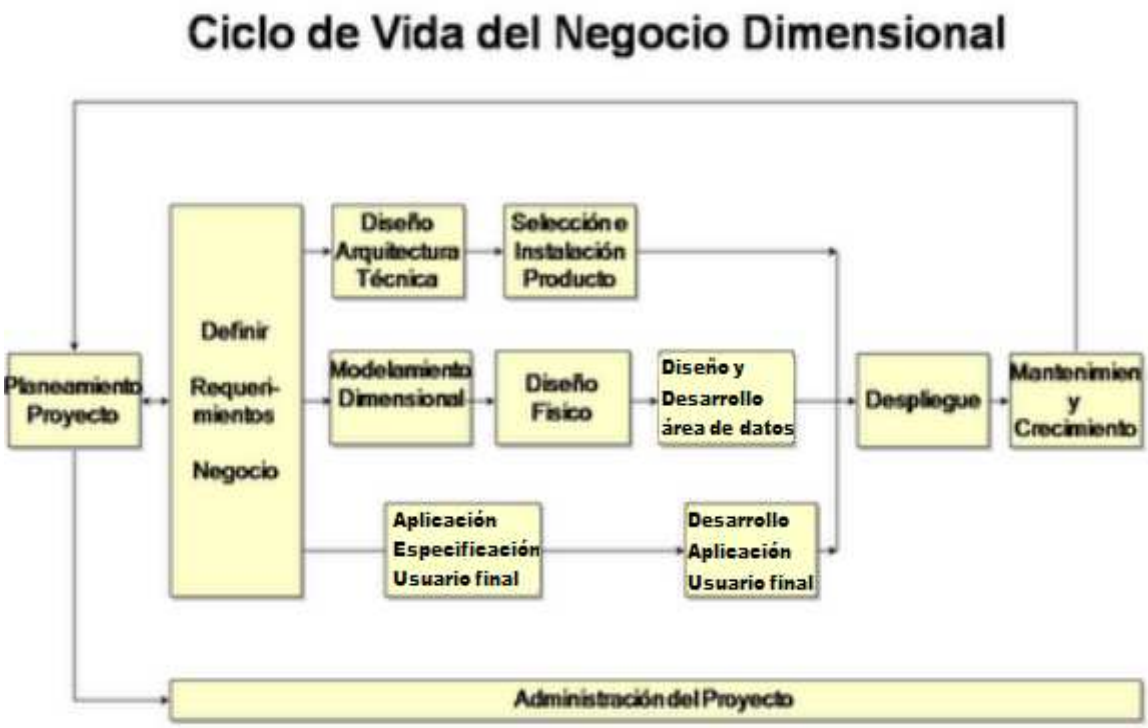

Figura 15: Ciclo de vida de la metodología de Ralph Kimball

4. Definición de los Requisitos

Una vez hecho el planteamiento del problema, describiremos de forma más concreta los requisitos exigidos para darle solución:

Identificador	F01	**Nombre**	Créditos troncales superados.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos troncales superados de los alumnos en su titulación.		

Identificador	F02	**Nombre**	Créditos obligatorios superados.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos obligatorios superados de los alumnos en su titulación.		

Identificador	F03	**Nombre**	Créditos optativos superados.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos optativos superados de los alumnos en su titulación.		

Identificador	F04	**Nombre**	Créditos de libre elección superados.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos de libre elección superados de los alumnos en su titulación.		

Identificador	F05	**Nombre**	Créditos troncales pendientes de superar.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos troncales pendientes de superar de los alumnos en su titulación.		

Identificador	F06	**Nombre**	Créditos obligatorios pendientes de superar.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos obligatorios pendientes de superar de los alumnos en su titulación.		

Identificador	F07	**Nombre**	Créditos optativos pendientes de superar.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos optativos pendientes de superar de los alumnos en su titulación.		

Identificador	F08	**Nombre**	Créditos de libre elección pendientes de superar.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer el número de créditos de libre elección pendientes de superar de los alumnos en su titulación.		

Identificador	F09	**Nombre**	Tiempo aprobado completo primer curso mayor de un año.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de primer curso mayor en un año a la fecha de matriculación de primer curso de su titulación.		

Identificador	F10	**Nombre**	Tiempo aprobado completo primer curso mayor de dos años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de primer curso mayor en dos años a la fecha de matriculación de primer curso de su titulación.		

Identificador	F11	**Nombre**	Tiempo de aprobado completo primer curso mayor de tres años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de primer curso mayor en tres años a la fecha de matriculación de primer curso de su titulación.		

Identificador	F12	**Nombre**	Tiempo de aprobado completo primer curso mayor de cuatros años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de primer curso mayor en cuatro años a la fecha de matriculación de primer curso de su titulación.		

Identificador	F13	**Nombre**	Tiempo de aprobado completo segundo curso mayor de un año.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de segundo curso mayor en un año a la fecha de matriculación de segundo curso de su titulación.		

Identificador	F14	**Nombre**	Tiempo de aprobado completo segundo curso mayor de dos años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de segundo curso mayor en dos años a la fecha de matriculación de segundo curso de su titulación.		

Identificador	F15	**Nombre**	Tiempo de aprobado completo segundo curso mayor de tres años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de segundo curso mayor en tres años a la fecha de matriculación de segundo curso de su titulación.		

Identificador	F16	**Nombre**	Tiempo de aprobado completo segundo curso mayor de cuatro años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de segundo curso mayor en cuatro años a la fecha de matriculación de segundo curso de su titulación.		

Identificador	F17	**Nombre**	Tiempo de aprobado completo tercer curso mayor de un año.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de tercer curso mayor en un año a la fecha de matriculación de tercer curso de su titulación.		

Identificador	F18	**Nombre**	Tiempo de aprobado completo tercer curso mayor de dos años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de tercer curso mayor en dos años a la fecha de matriculación de tercer curso de su titulación.		

Identificador	F19	**Nombre**	Tiempo de aprobado completo tercer curso mayor de tres años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de tercer curso mayor en tres años a la fecha de matriculación de tercer curso de su titulación.		

Identificador	F20	**Nombre**	Tiempo de aprobado completo tercer curso mayor de cuatro años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de tercer curso mayor en cuatro años a la fecha de matriculación de tercer curso de su titulación.		

Identificador	F21	**Nombre**	Tiempo de aprobado completo cuarto curso mayor de un año.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de cuarto curso mayor en un año a la fecha de matriculación de cuarto curso de su titulación.		

Identificador	F22	**Nombre**	Aprobado completo cuarto curso mayor de dos años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de cuarto curso mayor en dos años a la fecha de matriculación de cuarto curso de su titulación.		

Identificador	F23	**Nombre**	Tiempo de aprobado completo cuarto curso mayor de tres años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de cuarto curso mayor en tres años a la fecha de matriculación de cuarto curso de su titulación.		

Identificador	F24	**Nombre**	Tiempo de aprobado completo cuarto curso mayor de cuatro años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de cuarto curso mayor en cuatro años a la fecha de matriculación de cuarto curso de su titulación.		

Identificador	F25	**Nombre**	Tiempo de aprobado completo quinto curso mayor de un año.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de quinto curso mayor en un año a la fecha de matriculación de quinto curso de su titulación.		

Identificador	F26	**Nombre**	Tiempo de aprobado completo quinto curso mayor de dos años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de quinto curso mayor en dos años a la fecha de matriculación de quinto curso de su titulación.		

Identificador	F27	**Nombre**	Tiempo de aprobado completo quinto curso mayor de tres años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de quinto curso mayor en tres años a la fecha de matriculación de quinto curso de su titulación.		

Identificador	F28	**Nombre**	Tiempo de aprobado completo quinto curso mayor de cuatro años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de quinto curso mayor en cuatro años a la fecha de matriculación de quinto curso de su titulación.		

Identificador	F29	**Nombre**	Tiempo de aprobado completo sexto curso mayor de un año.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de sexto curso mayor en un año a la fecha de matriculación de sexto curso de su titulación.		

Identificador	F30	**Nombre**	Tiempo de aprobado completo sexto curso mayor de dos años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de sexto curso mayor en dos años a la fecha de matriculación de sexto curso de su titulación.		

Identificador	F31	**Nombre**	Tiempo de aprobado completo sexto curso mayor de tres años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de sexto curso mayor en tres años a la fecha de matriculación de sexto curso de su titulación.		

Identificador	F32	**Nombre**	Tiempo de aprobado completo sexto curso mayor de cuatro años.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con fecha de aprobado de sexto curso mayor en cuatro años a la fecha de matriculación de sexto curso de su titulación.		

Identificador	F33	**Nombre**	Nota media más alta de primer curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de primer curso más alta.		

Identificador	F34	**Nombre**	Nota media más alta de segundo curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de segundo curso más alta.		

Identificador	F35	**Nombre**	Nota media más alta de tercer curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de tercer curso más alta.		

Identificador	F36	**Nombre**	Nota media más alta de cuarto curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de cuarto curso más alta.		

Identificador	F37	**Nombre**	Nota media más alta de quinto curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de quinto curso más alta.		

Identificador	F38	**Nombre**	Nota media más alta de sexto curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de sexto curso más alta.		

Identificador	F39	**Nombre**	Nota media más baja de primer curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de primer curso más baja.		

Identificador	F40	**Nombre**	Nota media más baja de segundo curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de segundo curso más baja.		

Identificador	F41	**Nombre**	Nota media más baja de tercer curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de tercer curso más baja.		

Identificador	F42	**Nombre**	Nota media más baja de cuarto curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de cuarto curso más baja.		

Identificador	F43	**Nombre**	Nota media más baja de quinto curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de quinto curso más baja.		

Identificador	F44	**Nombre**	Nota media más baja de sexto curso.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer la titulación con la nota media de sexto curso más baja.		

Identificador	F45	**Nombre**	Años entre matriculación y finalización de la titulación.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Alta	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con mayor diferencia de tiempo entre la fecha de matriculación y la fecha de finalización de la titulación.		

Identificador	F46	**Nombre**	Años entre finalización de la titulación y finalización del PFC.
Tipo:	Funcional	**Fecha:**	23-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Alta	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con mayor diferencia de tiempo entre la fecha de finalización de la carrera y la fecha de finalización del Proyecto Fin de Carrera de la titulación.		

Identificador	F47	**Nombre**	Nota media más alta en asignaturas troncales de primer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas troncales superadas en primer curso.		

Identificador	F48	**Nombre**	Nota media más baja en asignaturas troncales de primer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas troncales superadas en primer curso.		

Identificador	F49	**Nombre**	Nota media más alta en asignaturas troncales de segundo curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas troncales superadas en segundo curso.		

Identificador	F50	**Nombre**	Nota media más baja en asignaturas troncales de segundo curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas troncales superadas en segundo curso.		

Identificador	F51	**Nombre**	Nota media más alta en asignaturas troncales de tercer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas troncales superadas en tercer curso.		

Identificador	F52	**Nombre**	Nota media más baja en asignaturas troncales de tercer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas troncales superadas en tercer curso.		

Identificador	F53	**Nombre**	Nota media más alta en asignaturas troncales de cuarto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas troncales superadas en cuarto curso.		

Identificador	F54	**Nombre**	Nota media más baja en asignaturas troncales de cuarto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas troncales superadas en cuarto curso.		

Identificador	F55	**Nombre**	Nota media más alta en asignaturas troncales de quinto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas troncales superadas en quinto curso.		

Identificador	F56	**Nombre**	Nota media más baja en asignaturas troncales de quinto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas troncales superadas en quinto curso.		

Identificador	F57	**Nombre**	Nota media más alta en asignaturas troncales de sexto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas troncales superadas en sexto curso.		

Identificador	F58	**Nombre**	Nota media más baja en asignaturas troncales de sexto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas troncales superadas en sexto curso.		

Identificador	F59	**Nombre**	Nota media más alta en asignaturas obligatorias de primer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas obligatorias superadas en primer curso.		

Identificador	F60	**Nombre**	Nota media más baja en asignaturas obligatorias de primer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas obligatorias superadas en primer curso.		

Identificador	F61	**Nombre**	Nota media más alta en asignaturas obligatorias de segundo curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas obligatorias superadas en segundo curso.		

Identificador	F62	**Nombre**	Nota media más baja en asignaturas obligatorias de segundo curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas obligatorias superadas en segundo curso.		

Identificador	F63	**Nombre**	Nota media más alta en asignaturas obligatorias de tercer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas obligatorias superadas en tercer curso.		

Identificador	F64	**Nombre**	Nota media más baja en asignaturas obligatorias de tercer curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas obligatorias superadas en tercer curso.		

Identificador	F65	**Nombre**	Nota media más alta en asignaturas obligatorias de cuarto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas obligatorias superadas en cuarto curso.		

Identificador	F66	**Nombre**	Nota media más baja en asignaturas obligatorias de cuarto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas obligatorias superadas en cuarto curso.		

Identificador	F67	**Nombre**	Nota media más alta en asignaturas obligatorias de quinto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas obligatorias superadas en quinto curso.		

Identificador	F68	**Nombre**	Nota media más baja en asignaturas obligatorias de quinto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas obligatorias superadas en quinto curso.		

Identificador	F69	**Nombre**	Nota media más alta en asignaturas obligatorias de sexto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas obligatorias superadas en sexto curso.		

Identificador	F70	**Nombre**	Nota media más baja en asignaturas obligatorias de sexto curso.
Tipo:	Funcional	**Fecha:**	25-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas obligatorias superadas en sexto curso.		

Identificador	F71	**Nombre**	Nota media más alta en asignaturas optativas de primer curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas optativas superadas en primer curso.		

Identificador	F72	**Nombre**	Nota media más baja en asignaturas optativas de primer curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas optativas superadas en primer curso.		

Identificador	F73	**Nombre**	Nota media más alta en asignaturas optativas de segundo curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas optativas superadas en segundo curso.		

Identificador	F74	**Nombre**	Nota media más baja en asignaturas optativas de segundo curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas optativas superadas en segundo curso.		

Identificador	F75	**Nombre**	Nota media más alta en asignaturas optativas de tercer curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas optativas superadas en tercer curso.		

Identificador	F76	**Nombre**	Nota media más baja en asignaturas optativas de tercer curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas optativas superadas en tercer curso.		

Identificador	F77	**Nombre**	Nota media más alta en asignaturas optativas de cuarto curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas optativas superadas en cuarto curso.		

Identificador	F78	**Nombre**	Nota media más baja en asignaturas optativas de cuarto curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas optativas superadas en cuarto curso.		

Identificador	F79	**Nombre**	Nota media más alta en asignaturas optativas de quinto curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas optativas superadas en quinto curso.		

Identificador	F80	**Nombre**	Nota media más baja en asignaturas optativas de quinto curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas optativas superadas en quinto curso.		

Identificador	F81	**Nombre**	Nota media más alta en asignaturas optativas de sexto curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de las asignaturas optativas superadas en sexto curso.		

Identificador	F82	**Nombre**	Nota media más baja en asignaturas optativas de sexto curso.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Media	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de las asignaturas optativas superadas en sexto curso.		

Identificador	F83	**Nombre**	Alumnos con media más alta de la titulación.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más alta de la titulación.		

Identificador	F84	**Nombre**	Alumnos con media más baja de la titulación.
Tipo:	Funcional	**Fecha:**	27-11-2009
Prioridad:	Alta	**Necesidad:**	Sí
Estabilidad:	Normal	**Verificable:**	Sí
Descripción:	El sistema deberá permitir conocer los alumnos con la nota media más baja de la titulación.		

5. Modelado Dimensional

En el proceso de diseño dimensional propuesto por Kimball se distinguen 5 etapas:

1. Definición del proceso de negocio
2. Definición del grano
3. Elección de las dimensiones
4. Identificación de los hechos
5. Detalle de las tablas de dimensión

5.1. Definición del proceso de negocio

En este primer paso seleccionamos los procesos de negocio del modelo. Entendemos como proceso de negocio cualquier actividad empresarial que se realiza en la organización y que normalmente cuenta con un sistema de recogida de datos. Realmente la mejor manera de seleccionar un proceso de negocio es escuchar a los usuarios. Cuando hablamos de proceso no nos referimos a un departamento de negocio de la organización, sino a un proceso productivo en sí mismo.

En el caso de estudio se quiere analizar las titulaciones con mejor nota media separadas por cursos, así como la nota media de la titulación en su conjunto. Además, se podrá conocer las notas medias más altas y más bajas de los diferentes cursos y más en concreto, las notas medias para las asignaturas troncales, obligatorias y optativas en los diferentes cursos.

En general, el DM permitirá analizar las notas medias de los alumnos y de los diferentes cursos académicos, separándolos por tipos de asignaturas y agrupándolos por las diferentes titulaciones a las que pertenece el alumnado. La información almacenada en el DM proporcionará el conocimiento indispensable para saber qué titulaciones y que cursos de éstas, son las que peor nota media tienen y por consiguiente, los cursos en los que el alumnado obtiene peores resultados. Para ello, es importante tener una visión completa del ciclo de vida de cada uno de los alumnos a lo largo de su paso por la universidad.

5.2. Definición del grano

Una vez que se ha definido el proceso de negocio, la siguiente tarea será la definición de la granularidad, o lo que es lo mismo, hasta qué nivel de detalle se quiere alcanzar en el modelo de DM y más concretamente en la tabla de hechos.

Lo más recomendable en la metodología de Kimball es desarrollar el modelo en torno a una granularidad baja obtenida a partir del proceso de negocio. Es decir, el objetivo es estructurar el modelo en torno a una información lo más detallada posible de tal manera que ésta no se pueda desglosar. La ventaja de estas informaciones básicas o atómicas es que ofrecen una gran flexibilidad en su análisis, y los datos en un modelo de

dimensión permiten las consultas directas por parte de los usuarios. Además, permiten responder a consultas que no podrían responderse con mayor granularidad.

Por otro lado, siempre es posible declarar un nivel de granularidad más alto para un proceso de negocio que representa una agregación de los datos atómicos. Sin embargo, esta opción también implica una limitación a la hora de detallar las dimensiones. Este aumento de la granularidad dará una mayor dificultad al usuario a la hora de profundizar en los detalles por lo que es recomendable dar el mayor detalle posible al grano.

Con estas perspectivas, el grano más apropiado para el propósito de este proyecto consistiría en asignar una fila por cada estudiante que agrupa el sistema. De manera que se pueda estructurar la información en torno a los datos que se puedan obtener del alumno, entre los que pueden encontrarse los siguientes:

- Fecha de admisión
- Fecha de matrícula
- Nota de acceso
- Carrera que estudia
- Área de la carrera que estudia
- Asignaturas de su carrera
- Fecha de matriculación en cada asignatura
- Fecha en la que aprueba cada asignatura
- Nota de cada asignatura aprobada
- Convocatoria en la que aprueba cada asignatura
- Fecha en la que supera cada curso
- Nota media de cada curso
- Nota media de la carrera
- etc.

Está claro que debido al interés en conocer la evolución de los alumnos a lo largo del tiempo dentro del sistema educativo universitario, la dimensión más importante que surge en el sistema es el tiempo. El tiempo engloba no solo la fecha en la que el alumno comienza sus estudios, sino que incluye diferentes hitos que se deben ir acumulando como parte de la información. Como resultado de ello podremos saber, por ejemplo, en qué momento se matricula de una asignatura de un determinado departamento o bien en qué momento se aprueba dicha asignatura.

Desde este punto de vista, es posible que los usuarios estén interesados en un análisis del alumnado a lo largo del tiempo con el fin de conocer sus evoluciones en las diferentes asignaturas del curso universitario. De esta manera se podrían obtener valiosos análisis en relación a qué cursos son los más difíciles para los estudiantes y por tanto, es posible que se pueda saber qué alumnos son los que abandonan los estudios en base a la dificultad que presenta cada curso.

El hecho de incluir una fila por alumno, permitirá al usuario utilizar el sistema desde el ámbito de las calificaciones de todos los alumnos de una determinada asignatura. De esta manera, el usuario puede conocer exactamente que asignaturas presentan las mayores tasas de suspensos y así, su dificultad.

Como se puede ver, esta granularidad ofrece al usuario un gran abanico de posibilidades a la hora de utilizar el DM que se pretende construir. Todo ello gracias a un grano fino, que nos permite combinar los datos de multitud de maneras y desde su concepción más básica ya que un DM casi siempre exige los datos expresados en el grano más bajo posible de cada dimensión, no porque se deseen consultas de muy bajo nivel sino por reducir las consultas a la obtención de detalles de forma precisa.

5.3. Elección de las dimensiones

Este paso plantea cómo describen los datos los usuarios del propio proceso de negocio. Con él, queremos incorporar a las tablas de hechos el conjunto de dimensiones, que representan los valores que asumen todas las posibles descripciones en cada contexto del proceso de negocio. Como regla general, si el tamaño del grano ha quedado suficientemente claro, generalmente las dimensiones se pueden identificar con bastante facilidad.

Una vez que se ha escogido el grano, analizamos qué dimensiones tiene asociadas. En este proyecto las dimensiones más apropiadas son la fecha, la asignatura, el alumno y la titulación.

Las tablas de dimensiones son catálogos de información complementaria necesaria para la presentación de los datos a los usuarios. Es decir, la información general complementaria a cada uno de los registros de la tabla de hechos.

En un modelo dimensional bien diseñado las tablas de dimensiones tienen muchas columnas o atributos. Estos atributos describen las filas de la tabla de dimensiones y habrá tantas descripciones significativas como sea posible. No es raro que una tabla de dimensiones tenga desde 50 hasta 100 atributos, por lo que las tablas de dimensión tienden a ser relativamente poco profundas en cuanto al número de filas pero se compensa con el gran número de columnas según la visión de Kimball de las tablas de dimensión expresada en [5].

Cada dimensión se define por su clave principal, designado por la notación de PK en la figura 16, que muestra la estructura de una de las dimensiones, la *asignatura*. Esta clave primaria sirve como base para la integridad referencial con la tabla de hechos a la que se une.

Para Kimball, las tablas de dimensión desempeñan un papel vital en el almacén de datos. Según este autor, el almacén de datos es tan bueno en tanto en cuanto los atributos de dimensión se escojan de manera correcta. El poder del almacén de datos es directamente proporcional a la calidad y profundidad de los atributos de cada dimensión. Cuanto más tiempo se destine a rellenar los valores de los atributos de una columna o a garantizar la calidad de los posibles valores de esa columna, mejor es el almacén de datos. Los mejores atributos de las tablas de dimensiones son los que miden valores discretos y deben consistir en palabras reales en lugar de abreviaturas crípticas. Por ejemplo en la dimensión *asignatura* se incluyen como atributos típicos una breve descripción (de 50 a 70 caracteres), un nombre de asignatura, el tipo de asignatura (troncal, obligatoria, optativa o libre elección), el número de créditos que tiene, el curso

al que pertenece, las horas de teoría y práctica que tiene la asignatura y el valor de la misma en cuanto a teoría y práctica.

Nombre Campo
IDAsignatura (PK)
Titulacion (FK)
Nombre
Descripción
Tipo
Créditos
Especialidad
Curso
Horas Teoría
Horas Práctica
Valor Teoría
Valor Práctica

Figura 16: Tabla de Dimensión Asignatura

En la figura 17 se pueden observar las dimensiones definidas para este proyecto. La dimensión *Fecha* proporcionará valores temporales. La dimensión *Asignatura* definirá cada una de las asignaturas que se encuentran dentro de las carreras universitarias contempladas en este trabajo. Por otro lado, la dimensión *Alumno* nos proporcionará la información del alumno a estudiar. Por último, la dimensión *Titulación* proporcionará a los usuarios la información de las titulaciones, por lo que, al igual que alumno, esta dimensión formará parte de otra como clave ajena. En este caso será clave ajena de asignatura, gracias a lo cual el usuario podrá advertir a que titulación pertenece una asignatura consultando la dimensión asignatura.

Figura 17: Tabla de hechos y dimensiones

Como se observa en la figura 17 que representa el esquema en estrella para este proyecto, cada una de las dimensiones tendrá una clave primaria que posteriormente aparecerá en la tabla de hechos en forma de clave ajena (Foreign key). Evidentemente la tabla de hechos tendrá una clave primaria, que en este caso, será el propio alumno puesto que al ser único cada alumno en la tabla de hechos, tal y como se definió en el paso 2 definición del grano, cada fila de la tabla de hechos se irá actualizando en el tiempo según vaya obteniendo logros académicos cada alumno.

5.4 Identificación de los hechos que poblarán cada fila de la tabla de hechos

Los hechos se usan para definir qué es lo que se quiere medir. Para Kimball, todos los hechos candidatos en un diseño deben ser fieles al grano definido en el paso 2 y los hechos que pertenezcan claramente a otro grano deben separarse en una tabla de hechos diferente.

Las medidas más útiles para incluir en una tabla de hechos son las *aditivas*, es decir, aquellas medidas que se pueden sumar, como por ejemplo la cantidad de producto vendido, los costes de producción o la calificación numérica media obtenida en una asignatura. Todas son medidas numéricas que pueden calcularse con la suma de varias cantidades de la tabla de hechos. En consecuencia, por lo general, Kimball aconseja que los *hechos* a almacenar en una tabla de hechos sean casi siempre valores numéricos, enteros o reales.

A menudo se describen los hechos como un valor permanente, principalmente como una guía para el diseñador para ayudarle a descifrar qué es un hecho frente a un atributo de dimensión. La calificación de una asignatura o los créditos superados por el alumno son ejemplos de ello porque pueden tener prácticamente cualquier valor dentro de un rango amplio. Por lo general, las tablas de hechos constituyen el 90 por ciento o más del espacio total consumido por una base de datos dimensional.

El objetivo ahora será determinar qué información aparecerá en la tabla de hechos. Partiendo de la definición del grano del segundo paso, y asumiendo que los hechos deben ser fieles a éste, es posible identificar por ejemplo los resultados de los alumnos para una cierta asignatura, las tasas de aprobados y suspensos o también es posible proporcionar la valoración numérica obtenida por el alumno, cuantificando dichos resultados según su valoración académica.

Los hechos se pueden considerar como las mediciones realizadas en la intersección de los valores clave de las diferentes dimensiones. Desde esta perspectiva, los hechos son la justificación de la tabla de hechos, y los valores clave son simplemente estructuras administrativas para identificar los hechos. Sin embargo, según el caso de estudio, los hechos no se gestionan de la misma manera en la propia tabla de hechos. Según Kimball, se pueden diferenciar hasta 3 formas de hacerlo:

- Transacciones
- Instantáneas Periódicas
- Instantáneas Acumulativas

Los tres tipos de medidas son las opciones más comunes para el grano de cualquier tabla de hechos. Los tres son útiles, y a menudo es necesario combinar dos de ellas para obtener una imagen completa de un proceso de negocio. En los tres casos las dimensiones serían las mismas en las tres tablas de hechos, sin embargo su administración es diferente.

- ***Transacciones:*** Las transacciones en la tabla de hechos representan una acción atómica que ocurre en un instante del tiempo determinado. A menudo no hay

ninguna garantía de que un registro exista para un determinado instante en la tabla de hechos de transacciones, ya que un registro sólo existe cuando se ha producido su transacción correspondiente. Por el contrario, no hay límite para el número de registros de un determinado hecho. La fecha en el registro de transacciones puede determinar un día concreto o puede contener una mayor precisión con las horas y minutos.

Normalmente los hechos representan cantidades y su valor depende de la clave de la transacción. Una vez insertada en la tabla, no suele volverse a revisar dicha transacción con el fin de hacer actualizaciones.

- ***Instantáneas Periódicas:*** Estas tablas de hechos representan un intervalo de tiempo predefinido. A diferencia de las transacciones, suelen existir registros para un determinado instante del tiempo. Por lo general, existe un único registro para cada combinación de las claves de las dimensiones importantes.

  Las instantáneas periódicas pueden tener cualquier número de hechos, dependiendo de qué medidas son posibles o útiles para calcular. Algunos de estos hechos serían extraordinariamente difíciles de calcular si se utilizaran las transacciones.

  A menudo la instantánea periódica es la única tabla que puede generar fácilmente una vista regular y predecible de las medidas importantes de una empresa, como pueden ser el cálculo de los ingresos y los costes.

- ***Acumulación de Instantáneas:*** La acumulación de instantáneas representa un período de tiempo con un comienzo y un fin específicos. La acumulación de instantáneas casi siempre contiene marcas de tiempo múltiples en las tablas de hechos. Algunas de ellas pueden tener que manejar valores nulos, porque habrá casos en los que dicha marca no se haya llegado a cumplir aún. En el caso de estudio de este proyecto podría darse esta circunstancia en las marcas de tiempo que determinan la finalización de los diferentes cursos académicos o las que marcan el aprobado en las asignaturas.

  Este tipo de tabla de hechos es atractivo cuando se realiza el seguimiento de elementos que tienen una vida limitada, con un inicio y un final, como por ejemplo coberturas de la póliza de seguros, asignaturas superadas en un periodo concreto de tiempo o alquileres de películas en un videoclub. La ventaja de la acumulación de instantáneas es que podemos obtener una gran cantidad de información útil sin restricciones sobre la dimensión de tiempo.

  De acuerdo con lo visto en las definiciones establecidas por Kimball parece claro que el tipo de proceso de negocio a estudiar en este proyecto encaja mejor en la acumulación de instantáneas que en los otros dos modelos planteados. Esto es debido a que se trata de un proceso con un tiempo establecido que comienza cuando el alumno se matricula y termina cuando abandona los estudios o finaliza la carrera. Por otro lado, este proceso se puede dividir fácilmente en una serie de hitos a realizar, como la superación de asignaturas y cursos, por lo que continuamente se irán realizando actualizaciones de la tabla de hechos. Sin embargo, la cuestión de las actualizaciones no se puede realizar con la misma facilidad en los otros dos modelos de la tabla de hechos.

Aplicando Acumulación de Instantáneas al estudio universitario: Si nos centramos en la dimensión fecha, podemos pensar en diferentes hitos o momentos que el usuario pudiera querer analizar. Esta es la forma de representar el ciclo de vida completo de una actividad o proceso, que tiene un principio y un final. Como podemos observar en la figura 18 las diferentes etapas del proceso de negocio se relacionan a modo de tubería abarcando desde que el alumno comienza su vida universitaria, hasta que finaliza o bien la abandona.

De esta manera, podemos pensar en el proceso universitario de un alumno como una tubería. El alumno va superando exámenes, asignaturas y cursos en un proceso continuo que solamente finaliza cuando ha aprobado todos los créditos exigidos en su titulación. Todos estos hitos dentro de la *tubería* son almacenados en el sistema a fin de poder realizar un análisis en función del tiempo y vida universitaria del alumno.

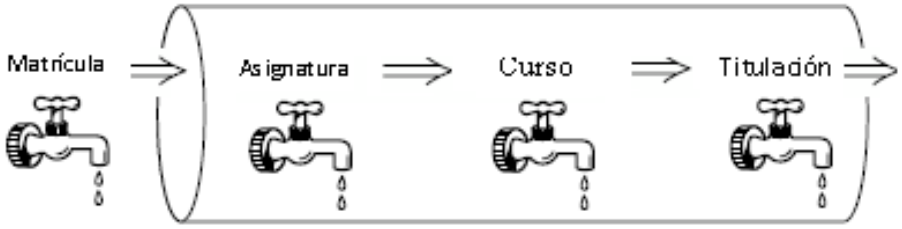

Figura 18: Diagrama de tubería de una carrera universitaria.

La acumulación de instantáneas nos permite ver el estado actualizado del alumno en cada momento, y en última instancia la posición del estudiante dentro del ciclo de vida universitario. Además, ofrece un gran número de alternativas en las que puede estar interesado el usuario como pueden ser asignaturas superadas en un determinado año, a que cursos pertenecen esas asignaturas, calificaciones de las mismas, créditos superados en el curso actual o en los anteriores. Es decir, esta forma de diseñar el proceso educativo nos ayudará a comprender mejor el estado en el que se encuentra cada alumno dentro de su carrera universitaria. Al establecer la acumulación de instantáneas como método para definir los hechos, nos damos cuenta de inmediato que el número de columnas de fecha es irremediablemente alto ya que se capturan un gran número de datos y fechas del estudiante progresando dentro de la *tubería*. Cada una de las fechas representa un hito importante en el cumplimiento de la misma.

La diferencia fundamental entre la acumulación de instantáneas y otro tipo de tablas de hechos es la idea de revisar y actualizar las filas de la tabla existente cuando la información está disponible. El grano de la tabla de hechos de la acumulación de instantáneas es una fila del menor nivel de detalle capturado. En nuestro caso, como se expuso anteriormente, el grano sería igual a una fila por estudiante, un grano que es fácil de identificar, ya que cada estudiante tiene sus propios datos personales que le hacen ser un sujeto único.

Resumiendo, podríamos concluir que la acumulación de instantáneas se ha elegido debido a que:

- Una sola fila representa la historia completa del alumno.

- Múltiples fechas representan los hitos de cada alumno.
- Los hechos con un comienzo y final acumulan sucesos interesantes para el proceso completo o *tubería*.
- Cada fila es revisada y modificada cada vez que algo sucede.

Una vez establecida la acumulación de instantáneas como método para introducir los hechos en la tabla de hechos, vamos a definir cuáles son los aspectos que nos permiten conseguir la mayor precisión posible en el proceso del negocio universitario que interesa en este proyecto. Teniendo en cuenta toda la información que se puede acumular para cada estudiante guardaremos lo siguiente en la tabla de hechos del proyecto.

Fecha de admisión (FK)
IDAsignatura (FK)
IDAlumno (PK)
IDTitulación (FK)
Fecha de solicitud de ingreso en la universidad (FK)
Nota de acceso en selectividad
Nota media del bachillerato
Nombre de la titulación en la que se matricula
Área en la que se matricula
Fecha de matriculación en primer curso (FK)
Fecha de aprobado completo del primer curso (FK)
Nota media de primer curso
Fecha de primera matriculación en segundo curso (FK)
Fecha de aprobado completo del segundo curso (FK)
Nota media de segundo curso
Fecha de primera matriculación en tercer curso (FK)
Fecha de aprobado completo del tercer curso (FK)
Nota media de tercer curso
Fecha de primera matriculación en cuarto curso (FK)
Fecha de aprobado completo del cuarto curso (FK)
Nota media de cuarto curso
Fecha de primera matriculación en quinto curso (FK)
Fecha de aprobado completo del quinto curso (FK)
Nota media de quinto curso
Fecha de primera matriculación en sexto curso (FK)
Fecha de aprobado completo del sexto curso (FK)
Nota media de sexto curso
Número de créditos troncales superados de primer curso
Número de créditos obligatorios superados de primer curso
Número de créditos optativos superados de primer curso
Número de créditos de libre elección superados de primer curso
Número de créditos troncales superados de segundo curso
Número de créditos obligatorios superados de segundo curso
Número de créditos optativos superados de segundo curso
Número de créditos de libre elección superados de segundo curso
Número de créditos troncales superados de tercer curso
Número de créditos obligatorios superados de tercer curso

Número de créditos optativos superados de tercer curso
Número de créditos de libre elección superados de tercer curso
Número de créditos troncales superados de cuarto curso
Número de créditos obligatorios superados de cuarto curso
Número de créditos optativos superados de cuarto curso
Número de créditos de libre elección superados de cuarto curso
Número de créditos troncales superados de quinto curso
Número de créditos obligatorios superados de quinto curso
Número de créditos optativos superados de quinto curso
Número de créditos de libre elección superados de quinto curso
Número de créditos troncales superados de sexto curso
Número de créditos obligatorios superados de sexto curso
Número de créditos optativos superados de sexto curso
Número de créditos de libre elección superados de sexto curso
Número de crédito troncales pendientes de superar de primer curso
Número de créditos obligatorios pendientes de superar de primer curso
Número de créditos optativos pendientes de superar de primer curso
Número de créditos de libre elección pendientes de superar de primer curso
Número de crédito troncales pendientes de superar de segundo curso
Número de créditos obligatorios pendientes de superar de segundo curso
Número de créditos optativos pendientes de superar de segundo curso
Número de créditos de libre elección pendientes de superar de segundo curso
Número de crédito troncales pendientes de superar de tercer curso
Número de créditos obligatorios pendientes de superar de tercer curso
Número de créditos optativos pendientes de superar de tercer curso
Número de créditos de libre elección pendientes de superar de tercer curso
Número de crédito troncales pendientes de superar de cuarto curso
Número de créditos obligatorios pendientes de superar de cuarto curso
Número de créditos optativos pendientes de superar de cuarto curso
Número de créditos de libre elección pendientes de superar de cuarto curso
Número de crédito troncales pendientes de superar de quinto curso
Número de créditos obligatorios pendientes de superar de quinto curso
Número de créditos optativos pendientes de superar de quinto curso
Número de créditos de libre elección pendientes de superar de quinto curso
Número de crédito troncales pendientes de superar de sexto curso
Número de créditos obligatorios pendientes de superar de sexto curso
Número de créditos optativos pendientes de superar de sexto curso
Número de créditos de libre elección pendientes de superar de sexto curso
Nota media de las asignaturas troncales superadas de primer curso
Nota media de las asignaturas obligatorias superadas de primer curso
Nota media de las asignaturas optativas superadas de primer curso
Nota media de las asignaturas de libre elección superadas de primer curso
Nota media de las asignaturas troncales superadas de segundo curso
Nota media de las asignaturas obligatorias superadas de segundo curso
Nota media de las asignaturas optativas superadas de segundo curso
Nota media de las asignaturas de libre elección superadas de segundo curso
Nota media de las asignaturas troncales superadas de tercer curso
Nota media de las asignaturas obligatorias superadas de tercer curso

Nota media de las asignaturas optativas superadas de tercer curso
Nota media de las asignaturas de libre elección superadas de tercer curso
Nota media de las asignaturas troncales superadas de cuarto curso
Nota media de las asignaturas obligatorias superadas de cuarto curso
Nota media de las asignaturas optativas superadas de cuarto curso
Nota media de las asignaturas de libre elección superadas de cuarto curso
Nota media de las asignaturas troncales superadas de quinto curso
Nota media de las asignaturas obligatorias superadas de quinto curso
Nota media de las asignaturas optativas superadas de quinto curso
Nota media de las asignaturas de libre elección superadas de quinto curso
Nota media de las asignaturas troncales superadas de sexto curso
Nota media de las asignaturas obligatorias superadas de sexto curso
Nota media de las asignaturas optativas superadas de sexto curso
Nota media de las asignaturas de libre elección superadas de sexto curso
Fecha de terminación de la titulación o carrera (FK)
Nota media de la carrera
Fecha de realización del Proyecto Final de Carrera (FK)
Nota Proyecto Final de Carrera

Figura 19: Tabla de hechos del proceso educativo universitario.

Como puede apreciarse en la tabla de hechos se ha incluido toda la información que se puede generar para un alumno a lo largo de su ciclo universitario. Los detalles sobre los tipos de asignaturas (troncales, obligatorias, optativas y de libre elección) permitirán conocer, por ejemplo, los motivos por los que un alumno tarda más de un año en terminar cada curso y averiguar si esa dilación se produce en un tipo de asignaturas o en otras.

5.5 Detalle de las tablas de dimensión

A continuación vamos a definir de manera clara y concisa cada una de las dimensiones que se han identificado para el proyecto y que ya se esbozaron en la figura 17, definiendo todos los atributos que componen las mismas.

5.5.1 Dimensión Fecha

Un parámetro que casi con toda probabilidad será común a todos los DM es el **tiempo**, ya que lo habitual es almacenar los hechos conforme van ocurriendo a lo largo del tiempo, obteniéndose así una serie temporal de la variable a estudiar. La dimensión fecha es la dimensión que se debe garantizar en todos los DM si queremos tener en cuenta series a lo largo del tiempo. De hecho, la fecha suele ser la primera dimensión en la ordenación de la base de datos de modo que la carga de los sucesivos intervalos de tiempo de los datos se coloca en una zona no utilizada del disco.

Dado que el tiempo es una dimensión presente en prácticamente cualquier DM merece una atención especial. Al diseñar la dimensión tiempo (tanto para un esquema en estrella como para un esquema en copo de nieve) hay que prestar especial cuidado, ya que puede hacerse de varias maneras y no todas son igualmente eficientes. La forma más común de diseñar esta tabla es poniendo como clave principal (PK) de la tabla la

fecha o **fecha/hora**. Este diseño no es de los más recomendables según Kimball, ya que a la mayoría de los sistemas de gestión de bases de datos les resulta más costoso hacer búsquedas sobre campos de tipo "**date**" o "**datetime**". Estos costes se reducen si el campo clave es de tipo **entero**, además, un dato entero siempre ocupa menos espacio que un dato de tipo fecha (el campo clave se puede repetir en millones de registros en la tabla de hechos y eso puede suponer mucho espacio), por lo que se mejorará el diseño de la tabla de fechas si se utiliza un campo "**TiempoID**" de tipo entero como clave principal.

A la hora de rellenar la tabla de fechas, si se ha optado por un campo de tipo entero para la clave, hay dos opciones, la que quizá sea más inmediata consiste en asignar valores numéricos consecutivos (*1, 2, 3, 4,*...) para los diferentes valores de fechas. La otra opción consistiría en asignar valores numéricos del tipo "***yyyymmdd***", es decir que los cuatro primeros dígitos del valor del campo indican el año de la fecha, los dos siguientes el mes y los dos últimos el día. Este segundo modo aporta una cierta ventaja sobre el anterior, ya que de esta forma se consigue que el dato numérico en sí, aporte por si solo la información de la fecha a la que se refiere, es decir, si en la tabla de hechos encontramos el valor ***20040723***, sabremos que se refiere al día 23 de julio de 2004; en cambio, con el primer método, podríamos encontrar valores como ***8456456*** y, para saber a qué fecha se refiere este valor tendríamos que hacer una consulta sobre otras columnas de la tabla de fechas.

Además del campo clave *TiempoID*, la tabla de dimensión fecha debe contener otros campos que también es importante estudiar. Entre los que Kimball recomienda estarían:

- Un campo "**año**".- Para contener valores como '2002', 2003, '2004', ...
- Un campo "**mes**".- Aquí se pueden poner los valores 'Enero', 'Febrero',... (o de forma abreviada: 'Ene', 'Feb',...).

Otro campo especial que se puede añadir es el "**Día de la semana**" ('lunes', 'martes',...), este campo se suele incluir cuando se quieren hacer estudios sobre el comportamiento de los días de la semana en general, por ejemplo es posible que un alumno tenga mejores resultados en exámenes que se realicen durante un sábado o quizás se le den mejor los que se celebren durante un día de diario.

En términos de usabilidad, el usuario de negocio típico no está versado en la semántica de las fechas que emplea el lenguaje SQL propio de los entornos de bases de datos, por lo que sería incapaz de aprovechar directamente las capacidades inherentes asociadas al tipo de datos fecha. Por otro lado, las funciones SQL de los campos fecha no admiten el filtrado por atributos, tales como los días entre semana frente a los fines de semana, días festivos, los períodos fiscales, las estaciones o acontecimientos importantes. A la vista de tales restricciones una tabla de dimensiones fecha explícita es esencial como aconseja Kimball.

Por último, a diferencia de la mayoría de otras dimensiones, la tabla de dimensión fecha se puede construir con antelación. Así podemos introducir en la tabla, filas para los días de los próximos 5 ó 10 años de manera que podamos cubrir toda la historia previa, así como varios años en el futuro. Incluso hablando de los días

equivalentes a 10 años, el tamaño de la tabla sería de alrededor de 3.650 líneas, lo que supone una tabla de dimensiones relativamente pequeña.

A continuación en la figura 20 se muestra la tabla fecha definida tal como se ha descrito:

TiempoID	Fecha	Año	Mes	Día de la semana	Cuatrimestre	Descripción de la fecha
20010922	22/02/2008	2008	Ene	'Lunes	1	22 de febrero de 2008
20010225	25/02/2009	2009	Feb	Sábado	1	25 de febrero de 2009
20020907	07/09/2007	2007	Feb	Martes	3	7de septiembre de 2007
20030127	27/01/2007	2001	Ene	Miércoles	1	21 de enero de 2007
20010222	14/06/2006	2006	Feb	Lunes	2	14 de junio de 2006

Figura 20: Atributos de la dimensión Fecha

El resultado final de la tabla dimensión fecha sería el reflejado en la figura 21 donde se muestra también la relación entre la dimensión fecha y la tabla de hechos.

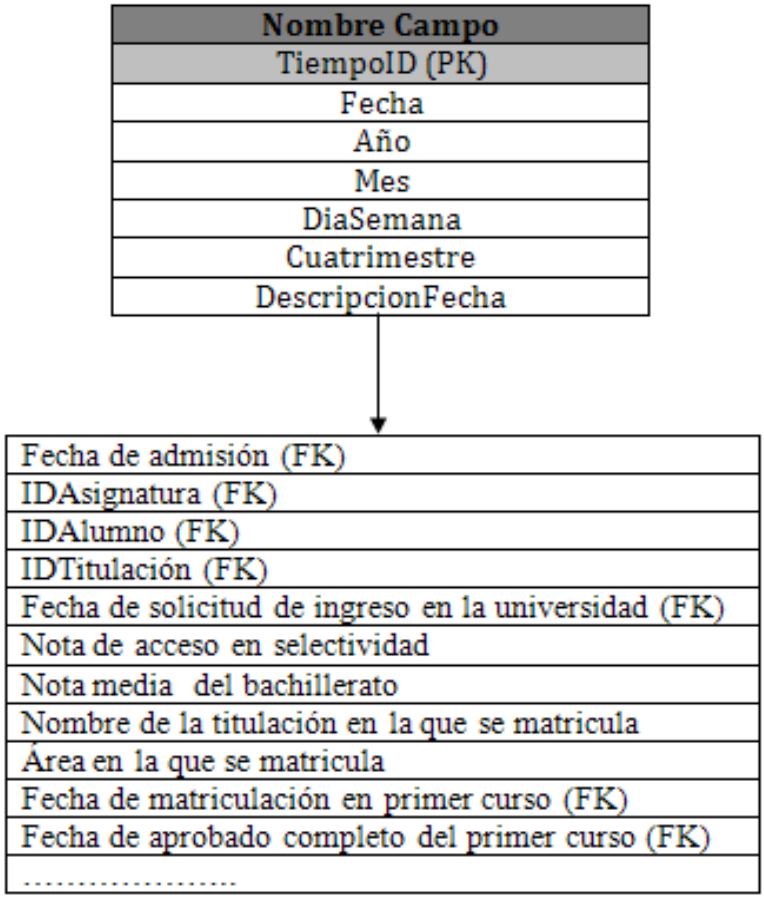

Figura 21: Dimensión fecha en el proceso educativo universitario.

5.5.2 Dimensión Asignatura

La dimensión asignatura describe cada una de las asignaturas que el alumno ha cursado o está pendiente de cursar.

Las carreras universitarias de las diferentes universidades tienen un número de créditos dispar, dependiendo de la escuela o facultad a la que pertenezcan y del número de años de los que está compuesta la titulación. Sin embargo, las carreras tienen algo en común, y es su división en asignaturas troncales, obligatorias, optativas y de libre elección. Sabiendo que la mayoría de asignaturas propuestas son de 6 créditos, esto hace un total de unas 50 asignaturas de media, de las cuales aproximadamente 15 son troncales, 15 son obligatorias, 15 son optativas y 5 son de libre elección. El resto de créditos corresponde a otros tipos de formación como pueden ser idiomas o cursos de humanidades. Ahora bien, si esas asignaturas las multiplicamos por el número de carreras que se imparten en la universidad tendríamos un total de unas 1200 ó 1400 asignaturas diferentes.

El problema que se presenta es el hecho de que no todos los alumnos estudian las mismas asignaturas optativas o de libre elección, y estas últimas no tienen que ser de la propia carrera que estudia el alumno, por lo que en la dimensión asignatura habría que incorporar atributos descriptivos que nos permitan identificar la carrera a la que pertenece en el caso de las asignaturas de libre elección o en el caso de las optativas, la especialidad. Esto podría resolverse mediante la introducción de un campo ***titulación*** que nos defina a qué carrera pertenece esa asignatura y que, como ya hemos comentado anteriormente, será clave ajena de asignatura debido a que la titulación es también una dimensión del proyecto. Teniendo en cuenta estos factores la dimensión asignatura sería la definida en la figura 22 en la que aparecen también algunos posibles valores**.**

AtignturaID	Nombre	Descripción	Tipo	Créditos	Especialidad
1	Física	Condensadores y comiente continua.	Troncal	7,5	
2	Algebra I	Métodos numéricos	Troncal	6	
3	Algebra II	Métodos numéricos avanzados	Troncal	6	
4	Redes de neuronas	Simulación aprendizaje.	Optativa	4,5	Inteligencia artificial
5	Microeconomía I	Oferta y demanda	Libre elección	4,5	

AsignaturaID	Curso	Horas Teoría	Horas Práctica	Valor Teoría	Valor Práctica	Titulación
1	1	5	0	100	0	Ing. Informática
2	1	4	0	100	0	Ing. Industrial
3	1	3	2	60	40	Ing. Industrial
4	3	1	3	25	75	Ing. Informática
5	4	2	0	100	0	Turismo

Figura 22: Dimensión asignatura.

El resultado final de la tabla dimensión asignatura sería el reflejado en la figura 23**.**

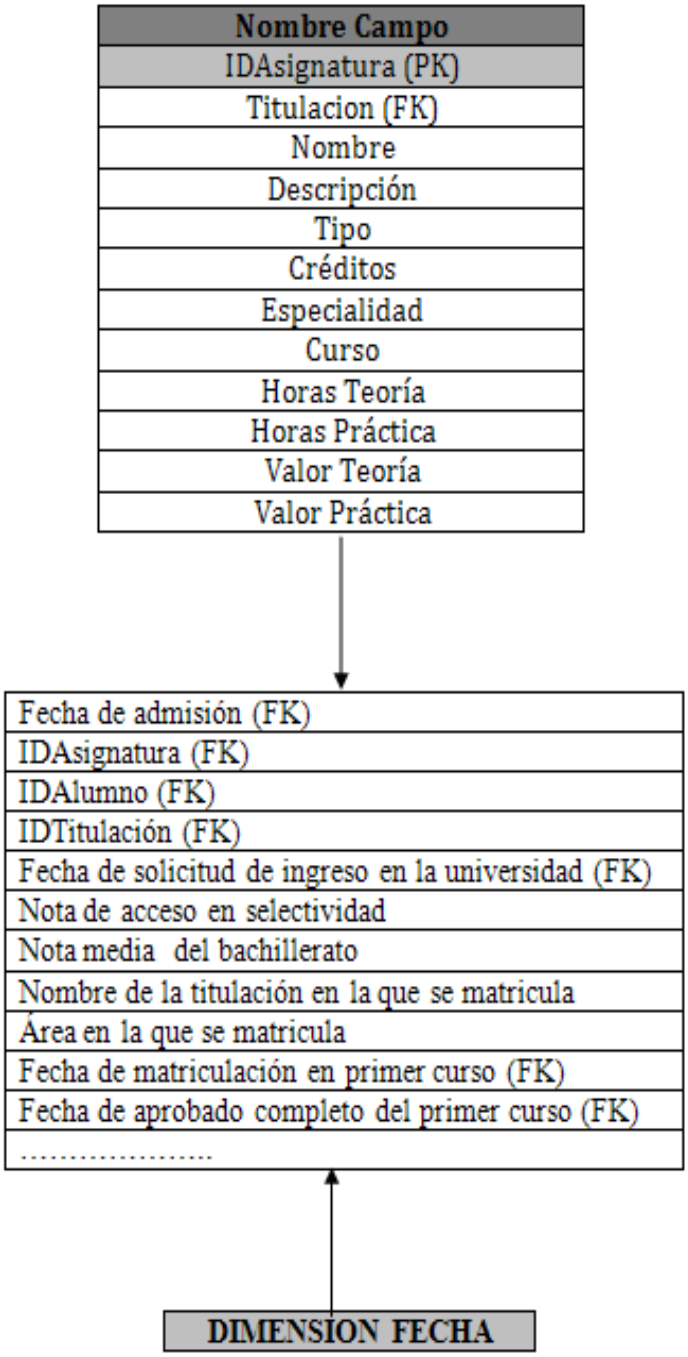

Figura 23: Dimensión Asignatura en el proceso educativo universitario.

5.5.3 Dimensión Alumno

La dimensión alumno debe dar identidad a cada uno de los alumnos que estudian en la universidad.

Quizás esta dimensión sea la más obvia de identificar sus atributos pero no por ello es la menos importante. Cada alumno vendrá identificado de manera inequívoca por el número de identificación dado por la propia universidad. Este número que se llamará NIA será único para cada alumno a lo largo del tiempo, incluso una vez que el alumno

haya abandonado los estudios universitarios, bien por haberlos concluido, bien por haberlos dejado incompletos.

Por otro lado, y de cara a posibles análisis que los usuarios del DM quieran realizar, se añadirán una serie de atributos con los que poder sesgar la información según lo requiera el usuario. La figura 24 muestra los atributos añadidos a la dimensión alumno con algunos valores de ejemplo:

IDAlumno	NIF	Nombre	Apellido1	Apellido2	Sexo	Fecha Nacimiento
1	00000000A	Luz	Casal	Ruiz	M	19901010
2	00000001B	Antonio	Rodríguez	Tames	V	19911111
3	00000002C	Víctor	Sandoval	Romero	V	19851110
4	00000003D	José	López	Martín	V	19881010
5	00000004E	Paloma	Sanz	Redondo	M	19860504

IDALumno	Provincia	Localidad	Dirección	E-mail	Instituto procedencia
1	Córdoba	Córdoba	C/ Cruces 14	Luz.casal@C3.es	Carlos III
2	Madrid	Madrid	C/Madrilejos 12	Ant.tames@C3.es	Siglo XXI
3	Madrid	Madrid	C/Universidad	Vic.romero@C3.es	Sagrada Familia
4	Madrid	Madrid	C/Salinas 12	Jose.lopez@C3.es	Sagrada Familia
5	Madrid	Madrid	C/De aguas 2	Pal.sanz@C3.es	Curro Romero

IDALumno	Nota acceso Universidad	Código Postal	Telefono1	Telefono2	País Nacimiento	Elección Estudios
1	6,8	28030	914897652	699598481	España	Bio Sanitaria
2	5.6	28009	915648231		España	Bio Sanitaria
3	9.4	28005	916648238	688415631	España	Bio Sanitaria
4	7.8	28410	918865142	647646213	España	Científico Técnica
5	6.5	28540	918764624	678156431	Francia	Científico Técnica

Figura 24: Atributos de la dimensión alumno.

Cada uno de los alumnos del sistema supondrá una fila en la tabla de hechos. Tal como se ha definido en la figura 24, podemos ver que un alumno quedará identificado de manera única mediante la clave primaria de la tabla, en este caso el ID del alumno, pero también mediante el número de identificación fiscal o NIF. Por otro lado, el sexo y la fecha de nacimiento, permitirán al usuario sesgar resultados en ciertas asignaturas o carreras por el género del individuo o la edad que tiene en el momento de la realización del examen. Los atributos instituto de secundaria y opción de estudios elegida (tecnológica, bio-sanitaria, humanidades o ciencias sociales) en el bachillerato también permitirán realizar interesante estudios relacionados con la procedencia de los alumnos.

En la siguiente figura añadimos la dimensión alumno al modelo en estrella.

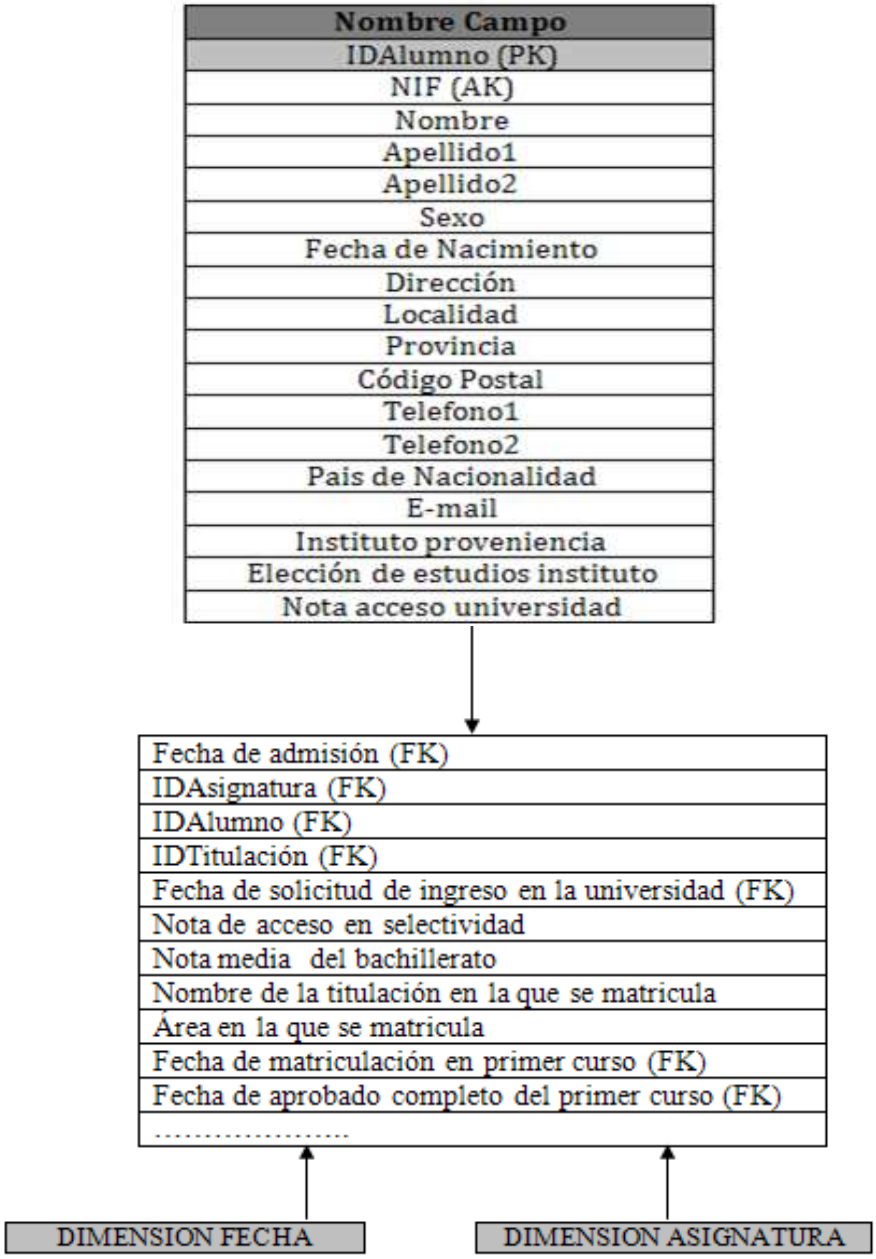

Figura 25: Dimensión Alumno en el proceso educativo universitario.

5.5.4 Dimensión Titulación

La dimensión titulación define la titulación a la que pertenecen las diferentes asignaturas contenidas en la dimensión asignatura.

Cada titulación vendrá definida por un ID que será común en los diferentes campus en donde se imparta. Además del ID tendrá asociado un nombre con el que se conoce a la titulación y los créditos troncales, obligatorios, optativos y de libre elección que se deben superar para obtener la titulación. Además, tendrá divididos por años los créditos que se deben cursar en cada año. Debido al proceso de implantación del

Espacio Europeo de Educación Superior que se lleva a cabo en las universidades españolas en el momento de desarrollar este proyecto, las titulaciones pueden ser antiguas o nuevas. En este último caso se denominan grados. Por este motivo, una titulación actualmente puede estar compuesta por tres (titulaciones antiguas de grado medio), cuatro (grados), cinco (titulaciones antiguas de grado superior y dobles titulaciones de grado) o seis (dobles titulaciones antiguas) cursos. Como la dimensión titulación incluye atributos para los seis cursos, en aquellas titulaciones donde no existan determinados cursos se permitirá la posibilidad de almacenar valores nulos

La estructura de la dimensión titulación se muestra en la **figura 26.**

IDTitulacion	Nombre	Créditos Troncales	Créditos obligatorios	Créditos optativos	Créditos de libre elección	Total créditos
1	Ing. Informática	122'5	120	95	37'5	375
2	Turismo	84	80	6	20	190
3	Ing. Industrial	122'5	120	95	37'5	375
4	Derecho+ Periodismo	180	180	100	45	505
5	Ing. Informática + Matemáticas	180	180	100	45	505

IDTitulacion	Créditos primer curso	Créditos segundo curso	Créditos tercer curso	Créditos cuarto curso	Créditos quinto curso	Créditos sexto curso
1	58'5	64'5	70'5	72	60	0
2	55	55	50	0	0	0
3	58'5	64'5	70'5	65	50	0
4	75	75	90	90	75	55
5	75	75	90	90	75	55

IDTitulacion	Director de la titulación
1	José Manuel Rodríguez
2	Sergio Sánchez Pérez
3	Javier Romero Cañas
4	Pedro Sanz García
5	José Luis Moreno Oro

Figura 26: Atributos de la dimensión titulación.

En la siguiente figura añadimos la dimensión titulación al modelo en estrella.

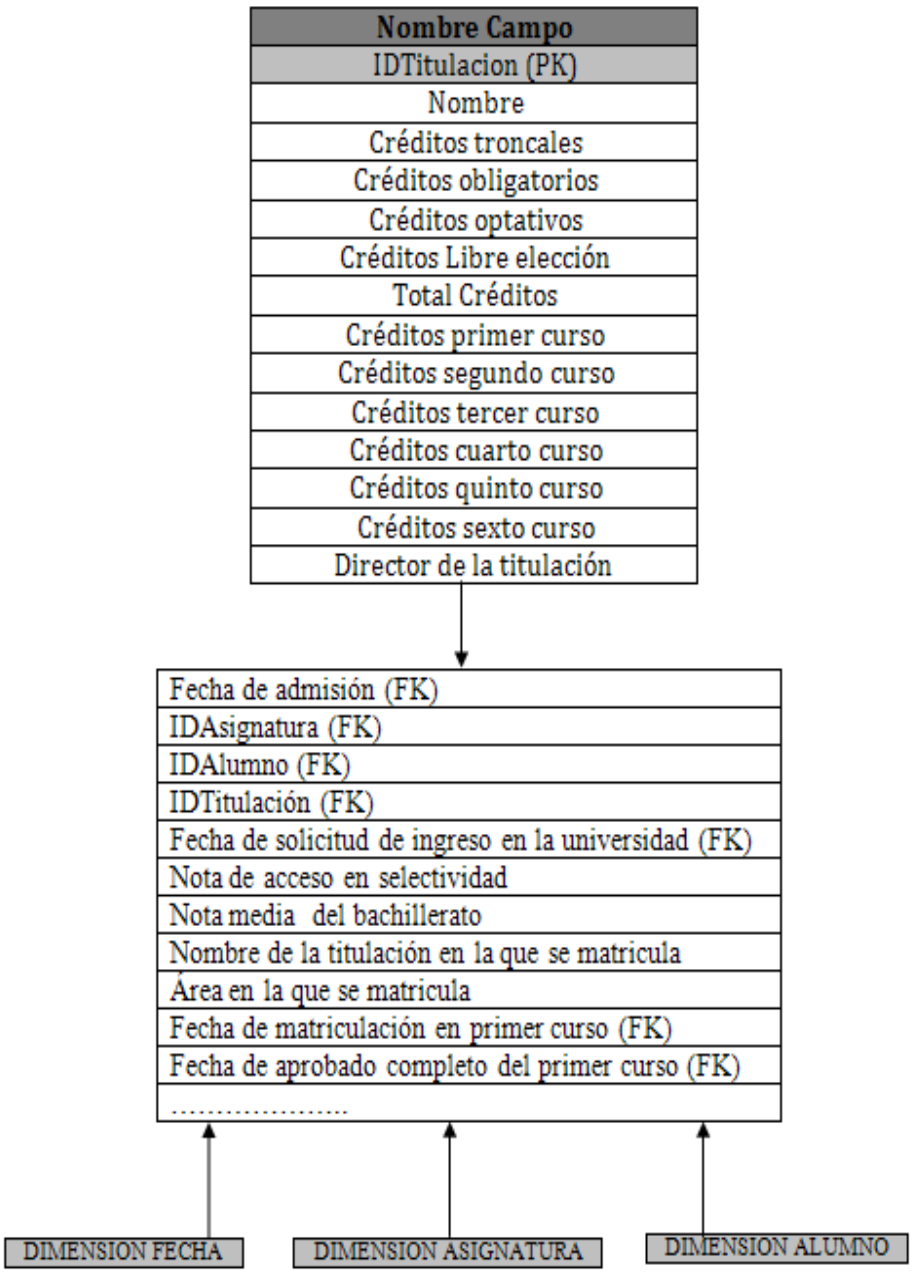

Figura 27: Dimensión Titulación en el proceso educativo universitario.

6. Diseño Físico del Data Warehouse

Siguiendo con la metodología de Ralph Kimball, el siguiente paso en el desarrollo del DM es el diseño físico del mismo. El modelo dimensional desarrollado en el capítulo anterior debe convertirse ahora en un diseño físico.

En el diseño físico se deben incluir los nombres de columna físicos, los tipos de los datos, las declaraciones de clave (si procede) y la posibilidad de incluir valores nulos. Contrariamente a la creencia general, añadir más hardware no es la única manera, ni la más eficaz, de ajustar el rendimiento. La creación de índices y tablas agregadas son alternativas mucho más rentables. En la figura 28 propuesta en [6] se pueden observar las diferencias entre los modelos lógico y físico.

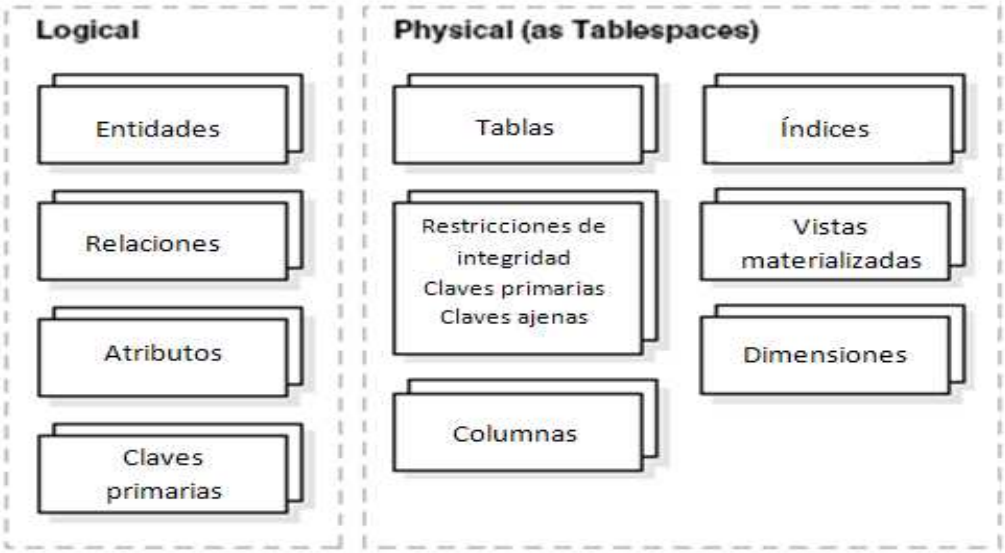

Figura 28: Comparación del modelo lógico y el físico.

En la figura 28 se pueden observar las diferencias entre los dos modelos. Mientras que en el modelo lógico se definen los objetos o entidades, sus atributos, sus claves primarias y las relaciones entre dichas entidades atendiendo a consideraciones lógicas, en el modelo físico se implementan estas estructuras de la manera más eficiente desde el punto de vista de la máquina.

Durante el proceso de diseño físico se trasladan los esquemas lógicos previstos a las estructuras reales. En este momento hay que transformar las entidades en tablas, crear las relaciones entre las dimensiones y la tabla de hechos mediante claves ajenas, transformar atributos en columnas y transformar los identificadores únicos primarios y alternativos en claves primarias y alternativas. Es de vital importancia decidir en este momento si se crearán tablas agregadas o índices para la optimización de las consultas en el sistema. Según se ha descrito en el punto 3.2.2.4. Diseño Físico, la creación de tablas agregadas está destinada a ofrecer una mejora en la optimización de consultas.

La siguiente figura muestra la descripción de la tabla de hechos desde el punto de vista físico:

Nombre del Campo	Tipo	Admite Nulos
Fecha_admisión (FK)	Fecha	NO
IDAsignatura (FK)	Numero	NO
IDAlumno (PK)	Numero	NO
IDTitulacion (FK)	Numero	NO
Fecha_solicitud_ingreso_universidad	Fecha	NO
Nota_acceso_selectividad	Numero	NO
Nota_media_bachillerato	Numero	NO
Titulacion_que_matricula	Varchar2(30)	NO
Area_que_matricula	Varchar2(30)	NO
Fecha_primera_matriculacion_primer_curso	Fecha	NO
Fecha_aprobado_completo_primer_curso	Fecha	SI
Nota_media_primer_curso	Fecha	SI
Fecha_primera_matriculacion_segundo_curso	Fecha	SI
Fecha_aprobado_completo_segundo_curso	Fecha	SI
Nota_media_segundo_curso	Numero	SI
Fecha_primera_matriculacion_tercer_curso	Fecha	SI
Fecha_aprobado_completo_tercer_curso	Fecha	SI
Nota_media_tercer_curso	Numero	SI
Fecha _primera_matriculacion_cuarto_curso	Fecha	SI
Fecha_aprobado_completo_cuarto_curso	Fecha	SI
Nota_media_cuarto_curso	Numero	SI
Fecha_primera_matriculacion_quinto_curso	Fecha	SI
Fecha_aprobado_completo_quinto_curso	Fecha	SI
Nota_media_quinto_curso	Numero	SI
Fecha_primera_matriculacion_sexto_curso	Fecha	SI
Fecha_aprobado_completo_sexto_curso	Fecha	SI
Nota_media_sexto_curso	Numero	SI
Num_creditos_troncales_superados_primer_curso	Numero	NO
Num_creditos_obligat_superados_primer_curso	Numero	NO
Num_creditos_optativ_superados_primer_curso	Numero	NO
Num_creditos_libre_elec_superados_primer_curso	Numero	NO
Num_creditos_troncales_superados_segundo_curso	Numero	SI
Num_creditos_obligat_superados_segundo_curso	Numero	SI
Num_creditos_optativ_superados_segundo_curso	Numero	SI
Num_creditos_libre_elec_superados_segundo_curso	Numero	SI
Num_creditos_troncales_superados_tercer_curso	Numero	SI
Num_creditos_obligat_superados_tercer_curso	Numero	SI
Num_creditos_optativ_superados_tercer_curso	Numero	SI
Num_creditos_libre_elec_superados_tercer curso	Numero	SI
Num_creditos_troncales_superados_cuarto_curso	Numero	SI
Num_creditos_obligat_superados_cuarto_curso	Numero	SI
Num_creditos_optativ_superados_cuarto_curso	Numero	SI
Num_creditos_libre_elec_superados_cuarto_curso	Numero	SI
Num_creditos_troncales_superados_quinto_curso	Numero	SI
Num_creditos_obligat_superados_quinto_curso	Numero	SI
Num_creditos_optativ_superados_quinto_curso	Numero	SI

Num_creditos_libre_elec_superados_quinto_curso	Numero	SI
Num_creditos_troncales_superados_sexto_curso	Numero	SI
Num_creditos_obligat_superados_sexto_curso	Numero	SI
Num_creditos_optativ_superados_sexto_curso	Numero	SI
Num_creditos_libre_elec_superados_sexto_curso	Numero	SI
Num_creditos_troncales_pendientes_superar_primer_curso	Numero	NO
Num_creditos_obligat_pendientes_superar_primer_curso	Numero	NO
Num_creditos_optativ_pendientes_superar_primer_curso	Numero	NO
Num_creditos_libre_elec_pendientes_superar_primer_curso	Numero	NO
Num_creditos_troncales_pendientes_superar_segundo_curso	Numero	SI
Num_creditos_obligat_pendientes_superar_segundo_curso	Numero	SI
Num_creditos_optativ_pendientes_superar_segundo_curso	Numero	SI
Num_creditos_libre_elec_pendientes_superar_segundo_curso	Numero	SI
Num_creditos_troncales_pendientes_superar_tercer_curso	Numero	SI
Num_creditos_obligat_pendientes_superar_tercer_curso	Numero	SI
Num_creditos_optativ_pendientes_superar_tercer_curso	Numero	SI
Num_creditos_libre_elec_pendientes_superar_tercer_curso	Numero	SI
Num_creditos_troncales_pendientes_superar_cuarto_curso	Numero	SI
Num_creditos_obligat_pendientes_superar_cuarto_curso	Numero	SI
Num_creditos_optativ_pendientes_superar_cuarto_curso	Numero	SI
Num_creditos_libre_elec_pendientes_superar_cuarto_curso	Numero	SI
Num_creditos_troncales_pendientes_superar_quinto_curso	Numero	SI
Num_creditos_obligat_pendientes_superar_quinto_curso	Numero	SI
Num_creditos_optativ_pendientes_superar_quinto_curso	Numero	SI
Num_creditos_libre_elec_pendientes_superar_quinto_curso	Numero	SI
Num_creditos_troncales_pendientes_superar_sexto_curso	Numero	SI
Num_creditos_obligat_pendientes_superar_sexto_curso	Numero	SI
Num_creditos_optativ_pendientes_superar_sexto_curso	Numero	SI
Num_creditos_libre_elec_pendientes_superar_sexto_curso	Numero	SI
Nota_media_asignaturas_troncales_superadas_primer_curso	Numero	SI
Nota_media_asignaturas_obligat_superadas_primer_curso	Numero	SI
Nota_media_asignaturas_optativ_superadas_primer_curso	Numero	SI
Nota_media_asignaturas_libre_elec_superadas_primer_curso	Numero	SI
Nota_media_asignaturas_troncales_superadas_segundo_curso	Numero	SI
Nota_media_asignaturas_obligat_superadas_segundo_curso	Numero	SI
Nota_media_asignaturas_optativ_superadas_segundo_curso	Numero	SI
Nota_media_asignaturas_libre_elec_superadas_segundo_curso	Numero	SI
Nota_media_asignaturas_troncales_superadas_tercer_curso	Numero	SI
Nota_media_asignaturas_obligat_superadas_tercer_curso	Numero	SI
Nota_media_asignaturas_optativ_superadas_tercer_curso	Numero	SI
Nota_media_asignaturas_libre_elec_superadas_tercer_curso	Numero	SI
Nota_media_asignaturas_troncales_superadas_cuarto_curso	Numero	SI
Nota_media_asignaturas_obligat_superadas_cuarto_curso	Numero	SI
Nota_media_asignaturas_optativ_superadas_cuarto_curso	Numero	SI
Nota_media_asignaturas_libre_elec_superadas_cuarto_curso	Numero	SI
Nota_media_asignaturas_troncales_superadas_quinto_curso	Numero	SI
Nota_media_asignaturas_obligat_superadas_quinto_curso	Numero	SI
Nota_media_asignaturas_optativ_superadas_quinto_curso	Numero	SI

Nota_media_asignaturas_libre_elec_superadas_quinto_curso	Numero	SI
Nota_media_asignaturas_troncales_superadas_sexto_curso	Numero	SI
Nota_media_asignaturas_obligat_superadas_sexto_curso	Numero	SI
Nota_media_asignaturas_optativ_superadas_sexto_curso	Numero	SI
Nota_media_asignaturas_libre_elec_superadas_sexto_curso	Numero	SI
Fecha_terminacion_titulacion	Fecha	SI
Nota_media_carrera	Numero	SI
Fecha_realizacion_Proyecto_Final_Carrera	Fecha	SI
Nota_Proyecto_Final_Carrera	Numero	SI

Figura 29: Diseño físico de la tabla de hechos.

A continuación se muestra el diseño físico de cada una de las dimensiones:

Dimensión Fecha:

Nombre Campo	Tipo	Admite Nulos
TiempoID (PK)	Numero	NO
Fecha	date	NO
Anyo	Numero	NO
Mes	Varchar2(3)	NO
DiaSemana	Varchar2(7)	NO
Cuatrimestre	Numero	NO
Descripcion_Fecha	Varchar2(25)	NO

Figura 30: Diseño físico de la dimensión fecha

Dimensión Asignatura:

Nombre Campo	Tipo	Admite Nulos
IDAsignatura (PK)	Numero	NO
Titulación (FK)	Número	NO
Nombre	Varchar2(30)	NO
Descripción	Varchar2(60)	NO
Tipo	Varchar2(15)	NO
Créditos	Numero	NO
Especialidad	Varchar2(20)	SI
Curso	Numero	NO
Horas_Teoria	Numero	SI
Horas_Practica	Numero	SI
Valor_Teoria	Numero	SI
Valor_Practica	Numero	SI

Figura 31: Diseño físico de la dimensión asignatura.

Dimensión Alumno:

Nombre Campo	Tipo	Admite Nulos
IDAlumno	Numero	NO
NIF	Varchar2(9)	NO
Nombre	Varchar2(15)	NO
Apellido1	Varchar2(30)	NO
Apellido2	Varchar2(30)	NO
Sexo	Varchar2(1)	NO
Fecha_Nacimiento	Date	NO
Direccion	Varchar2(40)	NO
Localidad	Varchar2(15)	NO
Provincia	Varchar2(15)	NO
Codigo_Postal	Numero	NO
Telefono1	Varchar2(9)	NO
Telefono2	Varchar2(9)	SI
Pais_Procedencia	Varchar2(15)	NO
email	Varchar2(40)	SI
Instituto_proveniencia	Varchar2(40)	NO
Eleccion_estudios_instituto	Varchar2(40)	NO
Nota_acceso_universidad	Numero	NO

Figura 32: Diseño físico de la dimensión alumno.

Dimensión Titulación:

Nombre Campo	Tipo	Admite Nulos
IDTitulacion	Numero	NO
Nombre	Varchar2(30)	NO
Creditos_troncales	Numero	NO
Creditos_obligatorios	Numero	NO
Creditos_optativos	Numero	NO
Creditos_libre_elec	Numero	NO
Creditos primer_curso	Numero	NO
Creditos_segundo_curso	Numero	NO
Creditos_tercer_curso	Numero	NO
Creditos_cuarto_curso	Numero	SI
Creditos_quinto_curso	Numero	SI
Creditos_sexto_curso	Numero	SI
Total_creditos	Numero	NO
Director_titulación	Varchar2(30)	NO

Figura 33: Diseño físico de la dimensión titulación.

Para completar el diseño físico del DM es necesario definir y establecer cada uno de los índices, tablas agregadas o ambos que son necesarios en el sistema para una buena optimización de las consultas que se quieran llevar a cabo. En función del tipo de análisis que se realice sobre los datos serán necesarias unas u otras optimizaciones sobre

el diseño inicial. Es importante saber que las tablas agregadas pueden tener un efecto muy significativo en el rendimiento de las consultas y en algunos casos se consigue incrementar el tiempo de resolución en un factor de 100 o incluso 1000, por lo que se puede suponer que no existen otros medios con ganancias tan espectaculares.

Observando el diseño físico actual, está claro que el interés del diseñador del DM es intentar abarcar la mayor cantidad posible de consultas posible con la información que se tiene en el diseño de la base de datos. Sobre este supuesto hay que trabajar para incorporar al diseño tablas agregadas que optimicen el sistema y cumplan ciertas pautas:

- Proporcionar un significativo incremento en el rendimiento de tantas categorías de consultas de los usuarios como sea posible.
- Ser completamente transparente a los usuarios finales y los diseñadores de aplicaciones a excepción de los evidentes beneficios de rendimiento.
- Beneficiar directamente a todos los usuarios del almacén de datos, independientemente de la herramienta de consulta que use cada uno.
- Impactar lo menos posible en el coste de la extracción de datos sobre el sistema.

Una vez que las tablas agregadas están definidas, el volcado de datos sobre las mismas se hace de igual manera que si fuera una tabla del diseño físico. Con estas premisas establecidas, algunas de las consultas que se pueden plantear para que los usuarios consulten el DM son las siguientes, divididas en bloques:

- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos troncales superados.*
- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos obligatorios superados.*
- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos optativos superados.*
- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos de libre elección superados.*

- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos troncales pendientes de superar.*
- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos obligatorios pendientes de superar.*
- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos optativos pendientes de superar.*
- *Por fecha de matriculación, titulación y alumno obtener el número total de créditos de libre elección pendientes de superar.*

- *Número de alumnos con fecha de aprobado del primer curso mayor en un año a la fecha de matriculación de primer curso, agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del primer curso mayor en dos años a la fecha de matriculación de primer curso agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del primer curso mayor en tres años a la fecha de matriculación de primer curso agrupados por titulación.*

- *Número de alumnos con fecha de aprobado del primer curso mayor en cuatro años a la fecha de matriculación de primer curso agrupados por titulación.*

- *Número de alumnos con fecha de aprobado del segundo curso mayor en un año a la fecha de matriculación de primer curso, agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del segundo curso mayor en dos años a la fecha de matriculación de primer curso agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del segundo curso mayor en tres años a la fecha de matriculación de primer curso agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del segundo curso mayor en cuatro años a la fecha de matriculación de primer curso agrupados por titulación.*

- *Número de alumnos con fecha de aprobado del tercer curso mayor en un año a la fecha de matriculación de primer curso, agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del tercer curso mayor en dos años a la fecha de matriculación de primer curso agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del tercer curso mayor en tres años a la fecha de matriculación de primer curso agrupados por titulación.*
- *Número de alumnos con fecha de aprobado del tercer curso mayor en cuatro años a la fecha de matriculación de primer curso agrupados por titulación.*

- *Número de alumnos con fecha de aprobado del cuarto curso mayor en un año a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay cuarto curso)*
- *Número de alumnos con fecha de aprobado del cuarto curso mayor en dos años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay cuarto curso)*
- *Número de alumnos con fecha de aprobado del cuarto curso mayor en tres años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay cuarto curso)*
- *Número de alumnos con fecha de aprobado del cuarto curso mayor en cuatro años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay cuarto curso)*

- *Número de alumnos con fecha de aprobado del quinto curso mayor en un año a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay quinto curso)*
- *Número de alumnos con fecha de aprobado del quinto curso mayor en dos años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay quinto curso)*
- *Número de alumnos con fecha de aprobado del quinto curso mayor en tres años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay quinto curso)*
- *Número de alumnos con fecha de aprobado del quinto curso mayor en cuatro años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay quinto curso)*
- *Número de alumnos con fecha de aprobado del sexto curso mayor en un año a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay sexto curso)*

- *Número de alumnos con fecha de aprobado del sexto curso mayor en dos años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay sexto curso)*
- *Número de alumnos con fecha de aprobado del sexto curso mayor en tres años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay sexto curso)*
- *Número de alumnos con fecha de aprobado del sexto curso mayor en cuatro años a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay sexto curso)*

- *Titulación con nota media de primer curso más alta.*
- *Titulación con nota media de segundo curso más alta.*
- *Titulación con nota media de tercer curso más alta.*
- *Titulación con nota media de cuarto curso más alta.*
- *Titulación con nota media de quinto curso más alta.*
- *Titulación con nota media de sexto curso más alta.*

- *Titulación con nota media de primer curso más baja.*
- *Titulación con nota media de segundo curso más baja.*
- *Titulación con nota media de tercer curso más baja.*
- *Titulación con nota media de cuarto curso más baja.*
- *Titulación con nota media de quinto curso más baja.*
- *Titulación con nota media de sexto curso más baja.*

- *Alumnos con la mayor diferencia entre la fecha de matriculación y la fecha de finalización de la carrera, agrupados por titulación.*
- *Alumnos con la mayor diferencia entre la fecha de finalización de la carrera y la fecha de finalización del Proyecto Fin de Carrera, agrupados por titulación.*

- *Alumnos con la nota media más alta de las asignaturas troncales superadas en primer curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas troncales superadas en segundo curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas troncales superadas en tercer curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas troncales superadas en cuarto curso agrupada por titulaciones. (si hay cuarto curso)*
- *Alumnos con la nota media más alta de las asignaturas troncales superadas en quinto curso agrupada por titulaciones. (si hay quinto curso)*
- *Alumnos con la nota media más alta de las asignaturas troncales superadas en sexto curso agrupada por titulaciones. (si hay sexto curso)*

- *Alumnos con la nota media más baja de las asignaturas troncales superadas en primer curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas troncales superadas en segundo curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas troncales superadas en tercer curso agrupada por titulaciones.*

- *Alumnos con la nota media más baja de las asignaturas troncales superadas en cuarto curso agrupada por titulaciones. (si hay cuarto curso)*
- *Alumnos con la nota media más baja de las asignaturas troncales superadas en quinto curso agrupada por titulaciones. (si hay quinto curso)*
- *Alumnos con la nota media más baja de las asignaturas troncales superadas en sexto curso agrupada por titulaciones. (si hay sexto curso)*

- *Alumnos con la nota media más alta de las asignaturas obligatorias superadas en primer curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas obligatorias superadas en segundo curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas obligatorias superadas en tercer curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas obligatorias superadas en cuarto curso agrupada por titulaciones. (si hay cuarto curso)*
- *Alumnos con la nota media más alta de las asignaturas obligatorias superadas en quinto curso agrupada por titulaciones. (si hay quinto curso)*
- *Alumnos con la nota media más alta de las asignaturas obligatorias superadas en sexto curso agrupada por titulaciones. (si hay sexto curso)*

- *Alumnos con la nota media más baja de las asignaturas obligatorias superadas en primer curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas obligatorias superadas en segundo curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas obligatorias superadas en tercer curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas obligatorias superadas en cuarto curso agrupada por titulaciones. (si hay cuarto curso)*
- *Alumnos con la nota media más baja de las asignaturas obligatorias superadas en quinto curso agrupada por titulaciones. (si hay quinto curso)*
- *Alumnos con la nota media más baja de las asignaturas obligatorias superadas en sexto curso agrupada por titulaciones. (si hay sexto curso)*

- *Alumnos con la nota media más alta de las asignaturas optativas superadas en primer curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas optativas superadas en segundo curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas optativas superadas en tercer curso agrupada por titulaciones.*
- *Alumnos con la nota media más alta de las asignaturas optativas superadas en cuarto curso agrupada por titulaciones. (si hay cuarto curso)*
- *Alumnos con la nota media más alta de las asignaturas optativas superadas en quinto curso agrupada por titulaciones. (si hay quinto curso)*
- *Alumnos con la nota media más alta de las asignaturas optativas superadas en sexto curso agrupada por titulaciones. (si hay sexto curso)*

- *Alumnos con la nota media más baja de las asignaturas optativas superadas en primer curso agrupada por titulaciones.*

- *Alumnos con la nota media más baja de las asignaturas optativas superadas en segundo curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas optativas superadas en tercer curso agrupada por titulaciones.*
- *Alumnos con la nota media más baja de las asignaturas optativas superadas en cuarto curso agrupada por titulaciones. (si hay cuarto curso)*
- *Alumnos con la nota media más baja de las asignaturas optativas superadas en quinto curso agrupada por titulaciones. (si hay quinto curso)*
- *Alumnos con la nota media más baja de las asignaturas optativas superadas en sexto curso agrupada por titulaciones. (si hay sexto curso)*

- *Alumnos con la nota media más alta en la carrera, agrupados por titulación.*
- *Alumnos con la nota media más baja en la carrera, agrupados por titulación.*

En función de las consultas expuestas anteriormente se van a crear diferentes tablas de agregación sobre la tabla de hechos de la siguiente manera:

TABLA DE AGREGACION 1
IDTitulacion (AK)
IDAlumno (PK)
Fecha_matriculacion_primer_curso
Num_creditos_troncales_superados_primer_curso
Num_creditos_obligat_superados_primer_curso
Num_creditos_optativ_superados_primer_curso
Num_creditos_libre_elec_superados_primer_curso
Num_creditos_troncales_superados_segundo_curso
Num_creditos_obligat_superados_segundo_curso
Num_creditos_optativ_superados_segundo_curso
Num_creditos_libre_elec_superados_segundo_curso
Num_creditos_troncales_superados_tercer_curso
Num_creditos_obligat_superados_tercer_curso
Num_creditos_optativ_superados_tercer_curso
Num_creditos_libre_elec_superados_tercer_curso
Num_creditos_troncales_superados_cuarto_curso
Num_creditos_obligat_superados_cuarto_curso
Num_creditos_optativ_superados_cuarto_curso
Num_creditos_libre_elec_superados_cuarto_curso
Num_creditos_troncales_superados_quinto_curso
Num_creditos_obligat_superados_quinto_curso
Num_creditos_optativ_superados_quinto_curso
Num_creditos_libre_elec_superados_quinto_curso
Num_creditos_troncales_superados_sexto_curso
Num_creditos_obligat_superados_sexto_curso
Num_creditos_optativ_superados_sexto_curso
Num_creditos_libre_elec_superados_sexto_curso

Tabla 1: Agregación sobre créditos superados.

En esta tabla se almacenarían todos los créditos superados por un alumno desde primer curso hasta el curso de la titulación que esté estudiando según lo definido en el primer bloque de consultas. Como se puede observar, la tabla agregada es una tabla más sobre la que es necesario definir una nueva clave primaria. Puesto que sobre esta tabla sigue siendo necesario el campo alumno, se volverá a crear la clave primaria sobre dicho campo. Además, para optimizar la búsqueda dentro de la tabla agregada, se crea un índice sobre el campo titulación para hacer la búsqueda aún más eficiente.

Para el segundo grupo de consultas establecidas, se crea otra tabla agregada que optimice las consultas al igual que la definida anteriormente pero esta vez sobre los créditos que los alumnos tienen aún pendientes de superar:

TABLA DE AGREGACION 2
IDTitulacion (AK)
IDAlumno (PK)
Fecha_matriculacion_primer_curso
Num_creditos_troncales_pendientes_superar_primer_curso
Num_creditos_obligat_pendientes_superar_primer_curso
Num_creditos_optativ_pendientes_superar_primer_curso
Num_creditos_libre_elec_pendientes_superar_primer_curso
Num_creditos_troncales_pendientes_superar_segundo_curso
Num_creditos_obligat_pendientes_superar_segundo_curso
Num_creditos_optativ_pendientes_superar_segundo_curso
Num_creditos_libre_elec_pendientes_superar_segundo_curso
Num_creditos_troncales_pendientes_superar_tercer_curso
Num_creditos_obligat_pendientes_superar_tercer_curso
Num_creditos_optativ_pendientes_superar_tercer_curso
Num_creditos_libre_elec_pendientes_superar_tercer_curso
Num_creditos_troncales_pendientes_superar_cuarto_curso
Num_creditos_obligat_pendientes_superar_cuarto_curso
Num_creditos_optativ_pendientes_superar_cuarto_curso
Num_creditos_libre_elec_pendientes_superar_cuarto_curso
Num_creditos_troncales_pendientes_superar_quinto_curso
Num_creditos_obligat_pendientes_superar_quinto_curso
Num_creditos_optativ_pendientes_superar_quinto_curso
Num_creditos_libre_elec_pendientes_superar_quinto_curso
Num_creditos_troncales_pendientes_superar_sexto_curso
Num_creditos_obligat_pendientes_superar_sexto_curso
Num_creditos_optativ_pendientes_superar_sexto_curso
Num_creditos_libre_elec_pendientes_superar_sexto_curso

Tabla 2: Agregación sobre créditos pendientes de superar.

Al igual que en la primera tabla de agregación, en esta tabla se crea un índice sobre el campo titulación para optimizar aún más las consultas. Para el tercer tipo de consultas definidas se crea la siguiente tabla de agregación:

TABLA DE AGREGACION 3
IDTitulacion (AK)
IDAlumno (PK)
Fecha_matriculacion_primer_curso
Fecha_aprobado_completo_primer_curso
Fecha_matriculacion_segundo_curso
Fecha_aprobado_completo_segundo_curso
Fecha_matriculacion_tercer_curso
Fecha_aprobado_completo_tercer_curso
Fecha_matriculacion_cuarto_curso
Fecha_aprobado_completo_cuarto_curso
Fecha_matriculacion_quinto_curso
Fecha_aprobado_completo_quinto_curso
Fecha_matriculacion_sexto_curso
Fecha_aprobado_completo_sexto_curso
Fecha_finalización_carrera
Fecha_finalización_Proyecto_Fin_Carrera

Tabla 3: Agregación de fecha matriculación y finalización de los cursos.

Para llevar a cabo las consultas relacionadas con las notas del alumnado dentro de las diferentes titulaciones se crea una última tabla agregada que contenga las notas de los diferentes tipos de asignaturas:

TABLA DE AGREGACION 4
IDTitulacion (AK)
IDAlumno (PK)
Nota_media_asignaturas_troncales_superadas_primer_curso
Nota_media_asignaturas_troncales_superadas_segundo_curso
Nota_media_asignaturas_troncales_superadas_tercer_curso
Nota_media_asignaturas_troncales_superadas_cuarto_curso
Nota_media_asignaturas_troncales_superadas_quinto_curso
Nota_media_asignaturas_troncales_superadas_sexto_curso
Nota_media_asignaturas_obligatorias_superadas_primer_curso
Nota_media_asignaturas_obligatorias_superadas_segundo_curso
Nota_media_asignaturas_obligatorias_superadas_tercer_curso
Nota_media_asignaturas_obligatorias_superadas_cuarto_curso
Nota_media_asignaturas_obligatorias_superadas_quinto_curso
Nota_media_asignaturas_obligatorias_superadas_sexto_curso
Nota_media_asignaturas_optativas_superadas_primer_curso
Nota_media_asignaturas_optativas_superadas_segundo_curso
Nota_media_asignaturas_optativas_superadas_tercer_curso
Nota_media_asignaturas_optativas_superadas_cuarto_curso
Nota_media_asignaturas_optativas_superadas_quinto_curso
Nota_media_asignaturas_optativas_superadas_sexto_curso

Tabla 4: Agregación sobre la nota media de asignaturas.

7 Diseño y Desarrollo de la Presentación de Datos

Este es el último paso en el desarrollo de los datos o la puesta en escena del sistema ETL. Como se describió en el punto 3.2.2.5. Diseño y Desarrollo de la Presentación de Datos, en esta etapa se lleva a cabo *la extracción, la transformación y la carga* (*ETL process*) de los datos dentro del diseño del DM. Además, se definen como *Procesos de Transformación* los procesos para convertir o recodificar los datos fuente a fin de poder efectuar la carga efectiva del Modelo Físico.

En este punto hay que decidir si comprar una herramienta para los procesos ETL o se va a desarrollar una por cuenta propia para llevar a cabo la tarea. En general, el autor de la metodología que se está siguiendo, Ralph Kimball, recomienda usar un producto comercialmente disponible, debido a que aunque no se recupere la inversión inicialmente a causa de la curva de aprendizaje, una herramienta proporcionará una mayor integración de metadatos y una mayor flexibilidad, reusabilidad, y facilidad de mantenimiento a largo plazo. Sin embargo, es importante valorar ambas alternativas antes de tomar una decisión teniendo en cuenta qué conocimiento de las herramientas tiene el equipo de trabajo y el presupuesto que se ha destinado para esta tarea.

Como el alcance de este proyecto está principalmente orientado al diseño, tanto lógico como físico de los datos del DM, no entraremos en profundidad en la realización de este apartado debido a que cualquiera de los dos casos, comprar o desarrollar una herramienta para llevar a cabo los procesos de ETL, están fuera de lo que se quiere llevar a cabo en el presente proyecto.

8 Diseño de la Arquitectura Técnica

Al igual que un plan para un nuevo hogar, la arquitectura técnica es el modelo de los servicios técnicos del DM y de sus elementos. El plan de la arquitectura técnica sirve como marco de organización para apoyar la integración de las tecnologías. La arquitectura permite detectar los problemas a priori y trata de minimizar al comienzo del proyecto las sorpresas que pudieran surgir.

En el capítulo 5 de la parte I, Arquitectura de un Data Warehouse, se mencionaron los componentes importantes de la arquitectura técnica, como el área de datos de organización, los servicios de acceso a datos y los metadatos. A continuación se dirigirá la atención a la arquitectura aplicable al proyecto de DM universitario.

Sobre las tres arquitecturas planteadas en el citado capítulo 5 se ha considerado más apropiado para este proyecto emplear la arquitectura con Área de Organización y Data Marts cuya figura reproducimos de nuevo por motivos de claridad y en la que queda representado el DM de gestión académica en el cual estamos trabajando.

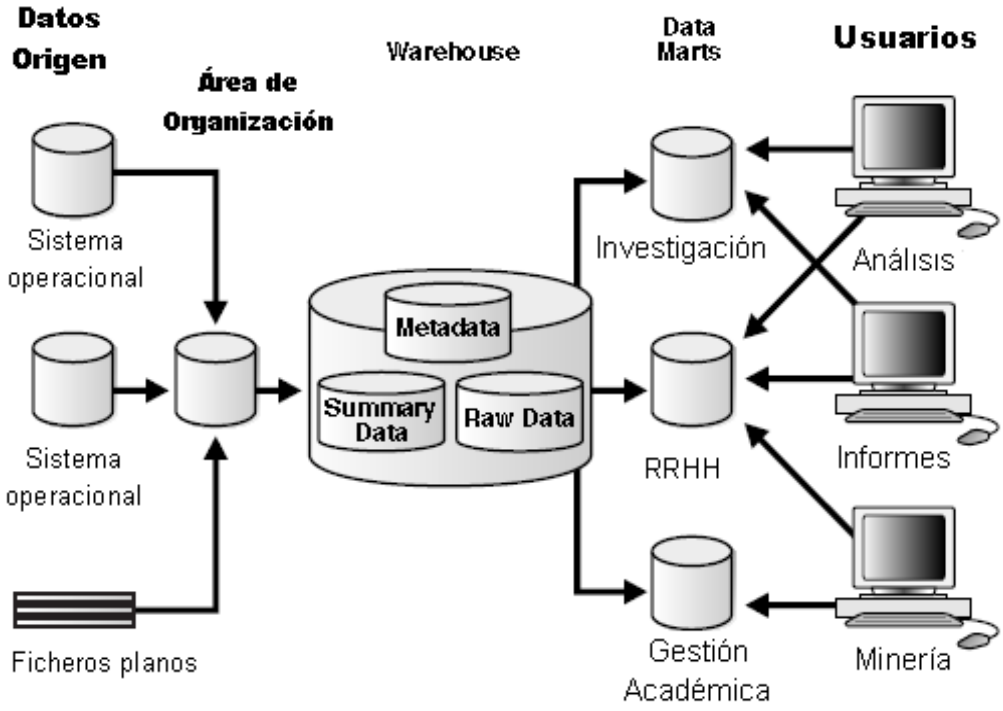

Figura 34: Arquitectura para el proyecto de DM se seguimiento académico de alumnos universitarios.

Se ha elegido esta arquitectura debido a que el área de organización o "*Staging Area*" es un área temporal donde se van a almacenar los datos necesarios de los sistemas origen. En este caso, se van a recogen los datos de las tablas alumno, asignatura y titulación de la base de datos operacional del entorno universitario que son los estrictamente necesarios para las cargas de datos en el DM. Una vez que los datos están cargados el DM se independiza de los sistemas origen hasta la siguiente carga. Lo único que se suele añadir es algún campo que almacene la fecha de la carga.

Obviamente estos datos no van a dar servicio a ninguna aplicación de *reporting*, son datos temporales que una vez hayan cumplido su función serán eliminados, de hecho en el esquema lógico de la arquitectura muchas veces no aparecen, ya que su función es meramente operativa. Las herramientas DSS o de *reporting* analítico accederán principalmente a los DM, pero también se pueden realizar consultas directamente sobre el DWH, sobre todo cuando sea necesario mostrar a la vez información que se encuentre en diferentes DM's. Dichas herramientas se analizarán con más detalles en los puntos sucesivos.

El último área de datos, es el lugar donde se crean los DM's. Éstos se obtienen a partir de la información recopilada en el área del DWH. Este área puede residir en la misma base de datos que las demás áreas de datos si la herramienta de explotación es de tipo ROLAP o también puede crearse ya fuera de la base de datos, en la estructura de datos propia que generan las aplicaciones de tipo MOLAP, más conocida como los cubos multidimensionales. En dicho área es donde residirá el DM desarrollado en el presente documento y que será consultado por los usuarios.

En la figura 34 se pueden observar diferentes DM dentro de la arquitectura de un DWH universitario y más concretamente se puede particularizar en el DM de "Gestión Académica" desarrollado en el presente documento. Dentro de la arquitectura de DWH para un entorno académico, también habría que destacar la creación de otros DM como pudieran ser los destinados al departamento de Recursos Humanos o el departamento de Investigación. Por otro lado los usuarios del sistema, análisis, extracción de informes y Datamining consultan los diferentes DM con el fin de obtener la información que desean.

9 Selección de Productos e Instalación

La instalación del producto sobre el que se implementará el desarrollo del DM viene condicionada por el coste que el producto supone a la empresa, en este caso la universidad que quiera llevar a cabo el desarrollo de un DM. Puesto que puede darse el caso de que una universidad ya disponga de licencia para usar alguna de las herramientas disponibles en el mercado, como por ejemplo *"Oracle Data Warehouse Builder"* de la empresa Oracle, el sistema podría implementarse sobre este producto, permitiendo un gran ahorro en cuanto a licencias.

El mayor coste para los desarrolladores vendrá determinado por el tiempo de aprendizaje del producto seleccionado, suponiendo este apartado un gasto apreciable para la empresa, sin embargo el coste ahorrado por la compra de licencias podría compensar con creces este gasto. Evidentemente, si se dispone ya de un software apropiado, la instalación no requiere tiempo y los desarrolladores pueden empezar directamente a conocer la herramienta.

Como se puede deducir, las decisiones sobre el producto a utilizar vienen en gran medida impuestas por el propio órgano encargado de compras y por la dirección y el departamento de informática que se reúnen para la toma de decisiones, buscando siempre la competitividad del producto a adquirir teniendo en cuenta el coste que se está dispuesto a soportar. Aunque no siempre es posible hacerlo, las decisiones deberían tomarse siempre teniendo en cuenta las ventajas tecnológicas y de rendimiento que ofrecen los productos comerciales.

A continuación se va a hacer un pequeño análisis de las principales herramientas comerciales disponibles en el mercado para intentar dilucidar qué ventajas y desventajas ofrecen unas frente a otras. Las herramientas analizadas son las ofrecidas por las compañías IBM, Oracle y Microsoft.

En cuanto a DB2 de la compañía IBM el producto *DB2 UDB Enterprise Server Edition* [10], es un sistema de gestión de bases de datos relacionales multiusuario altamente escalable que puede gestionar depósitos de datos, el proceso de análisis en línea (OLAP), el proceso de transacciones en línea (OLTP) y la minería de datos. También se pueden crear y gestionar entornos de bases de datos particionadas, las cuales pueden manejar grandes volúmenes de datos con adaptabilidad casi lineal, a la vez que se proporcionan ventajas como el aumento del rendimiento y una alta disponibilidad.

Otras de las herramientas más importantes del mercado es "Informix Metacube" [11]. Este producto ofrece a los administradores de las bases de datos una herramienta gráfica para manejar los metadatos que describen el almacén de datos de una manera lógica y amigable. De una manera sencilla, puede verse y modificarse el modelo de los DM sobre los cuáles los usuarios dependerán para acceder a los datos que componen el DWH. Además esta herramienta ofrece un tiempo de respuestas razonable y ofrece a los usuarios un mayor o menor detalle de los datos según lo precisen en cada momento.

Respecto a Oracle, aparte de las funcionalidades ofrecidas por su gestor de bases de datos tradicional, se destaca en el ámbito de los DWH el producto *Oracle Warehouse*

Builder [12]. Esta herramienta tiene una ventaja que radica en la gran variedad de funcionalidad que ofrece dentro de una sola herramienta. El modelado de datos y la calidad de los mismos son características centrales que cualquier herramienta para la integración de datos empresariales debe tener. No obstante, una ventaja estructural clave de *Oracle Warehouse Builder* es la integración de los componentes. *Oracle Warehouse Builder* proporciona todas sus capacidades dentro de un repositorio común y una interface de usuarios.

Por último, **Microsoft SQL Server 2005** [13] es un sistema para la gestión de bases de datos producido por Microsoft basado en el modelo relacional. Sus lenguajes para consultas son T-SQL (*Transact SQL*) y ANSI SQL.

Con SQL Server 2005, los Servicios de Análisis se mueven en el entorno del análisis en tiempo real. Los Servicios de Transformación de Datos son un conjunto de herramientas gráficas y objetos programables que pueden usarse para extraer, transformar y cargar datos (ETL) desde fuentes muy diversas y llevarlas a un destino único o múltiples destinos mediante un diseño que proporciona una plataforma ETL integral. Además sus servicios de Reporting permiten a las empresas integrar de forma sencilla datos desde fuentes heterogéneas y DWH en informes valiosos, interactivos y gestionables, que pueden localizarse y consultarse en intranets, extranets e Internet.

A continuación se van a mostrar unas tablas comparativas de las diferentes herramientas comentadas para valorar las ventajas y desventajas de usar unas u otras:

Sistemas operativos compatibles:

BBDD\SO	Windows	Mac OS X	Linux	UNIX
DB2	SI	SI	SI	SI
Microsoft SQL Server	SI	NO	NO	NO
Oracle	SI	SI	SI	SI
Informix-Metacube	SI	SI	SI	SI

Tabla 5: Compatibilidad de bases de datos con los sistemas operativos actuales.

Propiedades de Data Warehousing de las herramientas:

Criterio/Herramienta	MS SQL SERVER 7.0	Oracle	DB2	Informix-Metacube
Soporte de Base de datos/Data Warehouse	ALTO	ALTO	ALTO	ALTO
Soporte a ETL	ALTO	ALTO	MEDIO	MEDIO
Soporte a OLAP	ALTO	ALTO	MEDIO	ALTO
Soporte a DataMining	NULO	ALTO	ALTO	ALTO
Soporte Acceso a datos	BAJO	ALTO	BAJO	ALTO
Soporte a Reporting	NULO	MEDIO	NULO	MEDIO
Facilidad de uso	ALTO	BAJO	ALTO	ALTO
Facilidad de aprendizaje	ALTO	BAJO	MEDIO	MEDIO
Facilidad de Integración con otras herramientas	ALTO	ALTO	ALTO	ALTO
Coste de la herramienta	BAJO	ALTO	MEDIO	BAJO
Adaptabilidad al proyecto	ALTO	ALTO	ALTO	ALTO
Facilidad de administración y mantenimiento	ALTO	BAJO	ALTO	ALTO

Tabla 6: Propiedades Data Warehousing de las herramientas.

10 Especificación de Aplicaciones para Usuarios Finales

Las aplicaciones de usuario final proporcionan acceso a la mayoría de usuarios de negocio con el fin de generar informes con la información que quieran obtener. Son las interfaces a las que tiene acceso el usuario, al cual se le debe proveer de un mecanismo para que vea los datos a un alto nivel y que obtenga con ello la solución a preguntas específicas.

No entra en el alcance de este proyecto especificar con detalle las aplicaciones para los usuarios finales, aunque deberían tenerse presentes en este apartado todas las consideraciones de requisitos enunciadas en las etapas iniciales. Estos requisitos deberían ser la base de las especificaciones de las aplicaciones para usuarios finales, si bien siempre deberá permitirse un margen para las aportaciones que puedan añadir los usuarios finales en un instante como éste, en el que el DM podría estar ya operativo al tener la arquitectura técnica y el modelo de datos realizado. Esta facilidad de interactividad con el DM, aunque sea a nivel básico podría ofrecer nuevas perspectivas no contempladas al principio del proyecto. En otras ocasiones, las aplicaciones para usuarios finales se realizan en paralelo al diseño de la arquitectura técnica y/o el modelado conceptual y físico. La forma de trabajar en cualquier caso estará delimitada por el ciclo de vida y la metodología elegidas para el proyecto. En concreto, en la metodología de Kimball que se emplea en el desarrollo de este proyecto, se puede apreciar en las figuras 5 y 15 como las fases de especificación y desarrollo de aplicaciones para usuarios finales se pueden realizar en paralelo a las otras fases mencionadas, existiendo una realimentación entre todas las fases que redunda en un beneficio claro, al permitir añadir las aportaciones del usuario en un instante previo al despliegue del proyecto.

11 Desarrollo de Aplicaciones para Usuarios Finales

Por las razones descritas en el apartado anterior, el desarrollo de este punto pertenece única y exclusivamente al ámbito del usuario, por tanto, como el alcance de este proyecto se centra principalmente en la parte no visible para los usuarios, es decir, en las fases de modelado y construcción técnica del DM, no se entrará con detalle en esta fase, limitándonos a señalar que deberían implementarse aplicaciones a medida para los usuarios, que cumplan los requisitos establecidos en el apartado anterior, en aquellos casos en los que los usuarios puedan formalizar detalladamente sus funciones y modo de trabajo, mientras que para los usuarios que no puedan formalizar con detalle estas funciones quizá sería más recomendable adquirir alguna de las herramientas comerciales que suelen ser más flexibles en sus funciones.

12 Despliegue

Esta fase está orientada al correcto funcionamiento de las diferentes partes del DM. En ella deben coexistir la tecnología, los datos y las aplicaciones de usuarios finales con la máxima eficacia y eficiencia posible. Es esencial que exista un buen soporte técnico que tenga una gran capacidad de respuesta ante los retos que puedan surgir una vez que se ha instalado el sistema, en forma de incidencias o dudas que puedan tener los usuarios del DM.

Puesto que este punto se refiere a una fase en la que el sistema está completamente terminado, este proyecto no abarca de manera explícita este apartado.

13 Mantenimiento y crecimiento

Al igual que ocurría en el apartado anterior esta fase se refiere a la operativa del sistema una vez desplegado y, por tanto, este proyecto no abarca de manera explícita su desarrollo. La creación de un DWH es un proceso que acompaña a la evolución de la organización durante toda su historia. Se necesita continuar con las actualizaciones de forma constante para poder seguir la evolución de las metas por conseguir. Sin embargo, al contrario de los sistemas tradicionales, los cambios en el desarrollo deben ser vistos como signos de éxito. Es importante establecer unas prioridades para poder manejar los nuevos requerimientos de los usuarios y de esa forma poder evolucionar y crecer, de manera que en las semanas posteriores a la terminación del proyecto se debe tener en cuenta una cierta cercanía con el usuario, dado que se le debe prestar la mayor ayuda posible para que el proyecto comience a dar resultados lo más rápido posible.

14 Gestión del Proyecto

La Gestión del Proyecto tiene como actividad principal la planificación, el seguimiento y el control de las actividades y de los recursos humanos y materiales que intervienen en el desarrollo de un Sistema de Información. Una buena gestión de proyecto permite conocer en todo momento qué problemas se producen y resolverlos o paliarlos de manera inmediata.

El presente proyecto se planificó para llevarse a cabo en nueve meses, con una dedicación máxima de 132 horas al mes con un único ingeniero junior durante esos nueve meses y un ingeniero sénior dedicado a tiempo parcial durante cuatro de los nueve meses. Concluido el tiempo estimado, no se han producido desviaciones en el tiempo de entrega del proyecto así como desviaciones en el presupuesto dedicado a la finalización del mismo. A continuación se puede ver el presupuesto previsto para el proyecto desglosado según personal, material y gastos varios que se han incluido en el presupuesto.

UNIVERSIDAD CARLOS III DE MADRID
Escuela Politécnica Superior

PRESUPUESTO DE PROYECTO

1.- Autor:

Miguel Rodríguez Sanz

2.- Departamento:

Informática

3.- Descripción del Proyecto:

Definición y aplicación de metodologías de Data Warehousing en entornos académicos.

- Titulo: **ANÁLISIS Y DISEÑO DE UN DATA MART PARA EL SEGUIMIENTO ACADÉMICO DE**
- Duración (meses): **9**
- Tasa de costes Indirectos: **20%** **20%**

4.- Presupuesto total del Proyecto (valores en Euros):

50.000,00 Euros

5.- Desglose presupuestario (costes directos)

PERSONAL

Apellidos y nombre	N.I.F. (no rellenar - solo a titulo informativo)	Categoría	Dedicación (hombres mes) [a]	Coste hombre mes	Coste (Euro)	Firma de conformidad
Mingo Postiglioni, Jack Mario		Ingeniero Senior	4	4.289,54	17.158,16	
Rodríguez Sanz, Miguel		Ingeniero	9	2.694,39	24.249,51	
		Hombres mes 13		**Total**	41.407,67	

[a] 1 Hombre mes = 131,25 horas. Máximo anual de dedicación de 12 hombres mes (1575 horas)
Máximo anual para PDI de la Universidad Carlos III de Madrid de 8,8 hombres mes (1.155 horas)

EQUIPOS

Descripción	Coste (Euro)	% Uso dedicado proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable
PC Windows XP	500,00	100	9	60	75,00
PC Windows XP	500,00	30	9	60	22,50
					0,00
				Total	97,50

[d] Fórmula de cálculo de la Amortización:

$$\frac{A}{B} x C x D$$

A = nº de meses desde la fecha de facturación en que el equipo es utilizado
B = periodo de depreciación (60 meses)
C = coste del equipo (sin IVA)
D = % del uso que se dedica al proyecto (habitualmente 100%)

SUBCONTRATACIÓN DE TAREAS

Descripción	Empresa	Coste imputable
	Total	0,00

OTROS COSTES DIRECTOS DEL PROYECTO[e]

Descripción	Empresa	Costes imputable
Impresión documentos	WorkCenter	60,00
Desplazamiento		100,00
	Total	160,00

[e] Este capítulo de gastos incluye todos los gastos no contemplados en los conceptos anteriores, por ejemplo:

6.- Resumen de costes

Presupuesto Costes Totales	**Presupuesto Costes Totales**
Personal	41.408
Amortización	98
Subcontratación de tareas	0
Costes de funcionamiento	160
Costes Indirectos	8.333
Total	49.998

Figura 35: Presupuesto del proyecto

Conclusiones

El proyecto ha tratado de establecer un marco teórico que permitiera conocer con exactitud los fundamentos que sostiene la tecnología de DWH y DM. Partiendo de las más importantes metodologías que se ofrecen actualmente, se ha escogido la que mejor se adaptaba al proyecto y se ha desarrollado punto por punto para alcanzar los objetivos marcados. Una vez dentro del marco de trabajo que nos permitiera la implementación práctica de un DWH, hemos aplicado la metodología de Ralph Kimball al proyecto de seguimiento académico de alumnos en el entorno universitario, desarrollando un DM que nos ayudara a extraer información acerca del paso de los alumnos por su evaluación universitaria, haciendo especial énfasis en las fases de diseño de la arquitectura y el modelado de datos.

La implementación de un DM ha permitido obtener una mejor visión de los sucesos en el ámbito universitario así como dar un apoyo a la toma de decisiones. En un mundo competitivo como el actual, incluso dentro del ámbito universitario, no es posible quedarse al margen de la tecnología para apoyar la administración de los negocios. En este sentido, la construcción del DM juega un papel vital para la mejora y evolución de la universidad.

Para diseñar una buena arquitectura de DM y un modelado dimensional apropiado ha sido necesario conocer bien los requerimientos del negocio y hacer un estudio profundo de las fuentes externas que iban a suministrar los datos. Además, es preciso hacer un buen diseño del área de transformación de datos, aunque en este caso, se ha supuesto que las tablas de las que se nutre el DM tienen un diseño similar al empleado en las tablas de dimensión, lo que facilitaría bastante el trabajo de esta fase. Como se especificó en los objetivos, la implementación del modelo dimensional con sus tablas de hechos y de dimensiones ha supuesto el núcleo central del proyecto de desarrollo práctico de un DM de seguimiento académico de alumnos. Gracias a este modelado dimensional, el DM proporciona un esquema en el que se aprecia claramente cuáles son los componentes que lo forman y como se interrelacionan entre ellos, posibilitando una mayor flexibilidad para añadir nuevas fuentes de datos. En general, el DM representa una oportunidad de mejora dentro del ámbito universitario que queda reflejada en varios aspectos:

- *Mejora en la presentación de la información:* Ahora, la información tiene un formato más completo, consistente con los requerimientos de negocio y sobre todo accesible. Información útil y valiosa que irá creciendo y mejorando su valor a lo largo del tiempo.
- *Mejora en el proceso de negocio:* El DM provee una mejora en el proceso de toma de decisiones a la universidad con el fin de obtener mejores decisiones y más rápidas. Con ello se podrá obtener un mayor entendimiento entre las dos partes implicadas en el proceso de negocio: alumnos y personal docente. Como se recordará, uno de los objetivos del seguimiento académico es ofrecer posibles indicadores sobre los cursos, asignaturas y titulaciones con mayor nivel de fracaso académico, con el propósito de mejorar esos indicadores y ayudar a los alumnos a que terminen con éxito sus estudios.

Gracias al DM este tipo de conocimiento es más sencillo de conseguir, por lo que se prevé una mejora sustancial en este sentido.

El éxito final del proyecto dependerá en gran medida del uso y rendimiento que pueda ofrecer en el ámbito universitario, alcanzado su máximo valor cuando esté sumamente explotado, permitiendo al usuario (en este caso la universidad que implante el DM) obtener beneficios materiales con su uso, ya que no hay que olvidar que la justificación de una empresa para la inversión en la construcción de un DWH o un DM es que éstos proporcionen la necesidad de información estratégica que tiene valor para la empresa con independencia del tamaño que tenga. En este sentido, aunque este proyecto no aborda la implementación e implantación completa del sistema, deja muy avanzada su realización, de forma que únicamente faltaría la parte que consideramos el front-end del sistema, es decir, la parte directamente visible por los usuarios.

Futuras líneas

Este proyecto ha tratado de ofrecer un marco metodológico para el desarrollo de proyectos de Data Warehouse y/o Data Mart, así como una aplicación práctica de una metodología a un caso de estudio concreto. Sin embargo, existen al menos dos aspectos que no se han considerado en profundidad y que podrían ser objeto de un estudio detallado en posteriores trabajos:

a) En primer lugar, el proyecto ofrece respuestas completas a las fases de back-end, del sistema, es decir, a la parte más relacionada con la arquitectura técnica y el modelado de las bases de datos, sin embargo, la parte más visible a los usuarios o front-end solo se ha esbozado ligeramente por exceder los límites del proyecto. La implementación completa del proyecto pasaría por una elaboración precisa de esta fase.

b) En segundo lugar, a partir de la información almacenada en el DM de seguimiento académico se puede considerar otro de los aspectos fundamentales dentro del *"Business Intelligence"*. Nos referimos a la minería de datos. Existen en el mercado herramientas que la posibilitan con el fin de responder a preguntas que tradicionalmente llevan demasiado tiempo para poder ser resueltas. Estas herramientas permiten explorar las bases de datos en busca de patrones ocultos, tendencias y comportamientos, encontrando información predecible que un experto no puede llegar a encontrar fácilmente. Un sistema de minería de datos o Datamining debidamente acoplado a este sistema sobre los procesos educativos universitarios permitiría obtener multitud de datos interesantes de cara a mejoras dentro de dichos procesos.

Bibliografía y Artículos Consultados

- [1] William H. Inmon: Building the Data Warehouse, Technical Publishing Group, 1992
- [2] R. Kimball: The Data Warehouse Toolkit. Ed. John Wiley, 1996.
- [3] José A. Royo: Data Warehouse and DataMining, Departamento de Informática e Ingeniería de Sistemas, 2003
- [4] SAS Institute, Inc. SAS Rapid Warehousing Methodology, SAS Institute White Paper, 2001
- [5] R. Kimball: The Data Warehouse Lifecycle Toolkit. Ed. John Wiley, 1998.
- [6] Oracle Corporation. Data Warehousing Guide, December 2005
- [7] Claudio Casares: Tutorial Data Warehousing, 1999.
- [8] Connolly, T. y Begg C.: Sistemas de Bases de Datos, Un enfoque práctico para diseño, implementación y gestión. Pearson Addison Wesley, 2005.
- [9] Nader, J: Sistema de Apoyo Gerencial Universitario. Instituto Tecnológico de Buenos Aires, Argentina (2003).
- [10] IBM Corporation. IBM Informix MetaCube ROLAP Option, 2002.
- [11] Whei-Jen Chen, Angus Beaton, David Kline, Glen Johnson. DB2 UDB Evaluation Guide for Linux and Windows, RedBooks 2003
- [12] Oracle Corporation. Oracle Database 11g for DataWarehousing and Business Intelligence An Oracle, White Paper (2007)
- [13] Michelle Dumler. Microsoft SQL Server2005 Product Guide, 2005.

ANEXO 1: Script de creación de tablas.

```
DROP TABLE Fecha;
DROP TABLE Asignatura;
DROP TABLE Alumno;
DROP TABLE Titulacion;
DROP TABLE Seguim_Academico;

CREATE TABLE Fecha(
TiempoID                    varchar2(8),
fecha                       date,
Anyo                        number not null,
Mes                         number not null,
Dia_semana                  varchar2(10),
Cuatrimestre                number not null,
Descripcion_fecha           varchar2(20),
Primary key (TiempoID)
);


CREATE TABLE Asignatura(
AsignaturaID                number not null,
Titulacion                  number not null,
Nombre                      varchar2(30) not null,
Descripcion                 varchar2(60) not null,
Tipo                        varchar2(15) not null,
Creditos                    number not null,
Cuatrimestre                number not null,
Especialidad                varchar2(20) ,
Curso                       number not null,
Horas_teoria                number,
Horas_practica              number,
Valor_teoria                number,
Valor_Practica              number,
Primary key (AsignaturaID)
);

CREATE TABLE Alumno(
IDAlumno                    number,
NIF                         varchar2(9) not null,
Nombre                      varchar2(15) not null,
Apellido1                   varchar2(30) not null,
Apellido2                   varchar2(30) not null,
Sexo                        Varchar2(1) not null,
Fecha_Nacimiento            date not null,
Direccion                   varchar2(40) not null,
Localidad                   varchar2(15) not null,
Provincia                   varchar2(15) not null,
Codigo_Postal               number not null,
Telefono1                   varchar2(9) not null,
Telefono2                   varchar2(9),
```

```
Pais_Procedencia                varchar2(15) not null,
Email                           varchar2(40),
Instituto_Procedencia           varchar2(40) not null,
Eleccion_estudios               varchar2(40) not null,
Nota_acceso_universidad         number not null,
primary key (AlumnoID)
);
```

```
CREATE TABLE Titulacion(
IDTitulacion                    number,
Nombre                          varchar2(30) not null,
Creditos_Troncales              number not null,
Creditos_Obligatorios           number not null,
Creditos_Optativos              number not null,
Creditos_libre_eleccion         number not null,
Total_creditos                  number not null,
Creditos_primer_curso           number not null,
Creditos_segundo_curso          number not null,
Creditos_tercer_curso           number not null,
Creditos_cuarto_curso           number,
Creditos_quinto_curso           number,
Creditos_sexto_curso            number,
Director_titulación             varchar2(30) not null,
primary key (IDTitulacion)
);
```

```
CREATE TABLE Seguim_Academico (
Fecha_admision                  date not null,
IDAsignatura                    number not null,
IDAlumno                        number not null,
IDTitulacion                    number not null,
Num_identificacion_alumno       number not null,
Fec_solic_ingreso_univ          date not null,
Nota_acceso_selectividad        number not null,
Nota_acceso_bachillerato        number not null,
Titulacion_que_matricula        Varchar2(30) not null,
Area_que_matricula              Varchar2(30) not null,
Fec_aprob_complet_1er_curso     date,
Fec_matriculacion_1er_curso     date not null,
Nota_media_primer_curso         date,
Fec_matriculacion_2_curso       date,
Fec_aprob_complet_2_curso       date,
Nota_media_segundo_curso        number,
Fec_matriculacion_3er_curso     date,
Fec_aprob_complet_3er_curso     date,
Nota_media_3er_curso            number,
Fec_matriculacion_4_curso       date,
```

```
Fec_aprob_complet_4_curso          date,
Nota_media_4_curso                 number,
Fec_matriculacion_5_curso          date,
Fec_aprob_complet_5_curso          date,
Nota_media_5_curso                 number,
Fec_matriculacion_6_curso          date,
Fec_aprob_complet_6_curso          date,
Nota_media_6_curso                 number,
cred_tronc_supera_1er_curso        number not null,
cred_oblig_supera_1er_curso        number not null,
cred_optat_supera_1er_curso        number not null,
cred_LE_supera_1er_curso           number not null,
cred_tronc_supera_2_curso          number,
cred_obligat_superados_2_curso     number,
cred_optativ_superados_2_curso     number,
cred_LE_superados_2_curso          number,
cred_tronc_superados_3er_curso     number,
cred_oblig_superados_3er_curso     number,
cred_optat_superados_3er_curso     number,
cred_LE_superados_3er_curso        number,
cred_tronc_superados_4_curso       number,
cred_oblig_superados_4_curso       number,
cred_optat_superados_4_curso       number,
cred_LE_superados_4_curso          number,
cred_tronc_supera_5_curso          number,
cred_oblig_supera_5_curso          number,
cred_optativ_superados_5_curso     number,
cred_LE_superados_5_curso          number,
cred_tronc_superados_6_curso       number,
cred_oblig_superados_6_curso       number,
cred_optat_superados_6_curso       number,
cred_LE_superados_6_curso          number,
cred_tronc_pendientes_1er_cur      number not null,
cred_oblig_pendientes_1er_cur      number not null,
cred_optat_pendientes_1er_cur      number not null,
cred_LE_pendientes_1er_curso       number not null,
cred_tronc_pendientes_2_cur        number,
cred_oblig_pendientes_2_cur        number,
cred_optat_pendientes_2_cur        number,
cred_LE_pendientes_2_curso         number,
cred_tronc_pendientes_3er_cur      number,
cred_oblig_pendientes_3er_cur      number,
cred_optat_pendientes_3er_cur      number,
cred_LE_pendientes_3er_curso       number,
cred_tronc_pendientes_4_cur        number,
cred_oblig_pendientes_4_cur        number,
cred_optat_pendientes_4_cur        number,
cred_LE_pendientes_4_curso         number,
```

```
cred_tronc_pendientes_5_curso        number,
cred_oblig_pendientes_5_curso        number,
cred_optat_pendientes_5_curso        number,
cred_LE_pendientes_5_curso           number,
cred_tronc_pendientes_6_curso        number,
cred_oblig_pendientes_6_curso        number,
cred_optat_pendientes_6_curso        number,
cred_LE_pendientes_6_curso           number,
media_asig_tronc_supera_1o           number,
media_asig_oblig_supera_1o           number,
media_asig_optat_supera_1o           number,
media_asig_LE_supera_1o              number,
media_asig_tronc_supera_2o           number,
media_asig_oblig_supera_2o           number,
media_asig_optat_supera_2o           number,
media_asig_LE_supera_2o              number,
media_asig_tronc_superdas_3o         number,
media_asig_oblig_superadas_3o        number,
media_asig_optat_superadas_3o        number,
media_asig_LE_superadas_3o           number,
media_asig_tronc_superadas_4o        number,
media_asig_oblig_superadas_4o        number,
media_asig_optat_superadas_4o        number,
media_asig_LE_superadas_4o           number,
media_asig_tronc_superadas_5º        number,
media_asig_oblig_superadas_5o        number,
media_asig_optat_superadas_5o        number,
media_asig_LE_superadas_5o           number,
media_asig_tronc_superadas_6o        number,
media_asig_oblig_superadas_6o        number,
media_asig_optat_superadas_6o        number,
media_asig_LE_superadas_6o           number,
Fecha_terminacion_titulacion         date,
Nota_media_carrera                   number,
Fecha_realizacion_PFC                date,
Nota_Proyecto_Final_Carrera          number,
primary key (IDAlumno)
);
```

```
ALTER TABLE ASIGNATURA
  ADD CONSTRAIT IDTITULACION
  FOREIGN KEY (Titulacion)
  REFERENCES TITULACION (IDTitulacion);
```

```
ALTER TABLE Seguim_Academico
  ADD CONSTRAIT IDTITULACION
  FOREIGN KEY (IDTitulacion)
  REFERENCES TITULACION (IDTitulacion);


ALTER TABLE Seguim_Academico
  ADD CONSTRAIT FECHA
  FOREIGN KEY (TiempoID)
  REFERENCES FECHA (TiempoID);


ALTER TABLE Seguim_Academico
  ADD CONSTRAIT ALUMNO
  FOREIGN KEY (IDAlumno)
  REFERENCES ALUMNO (IDAlumno);


ALTER TABLE Seguim_Academico
  ADD CONSTRAIT ASIGNATURA
  FOREIGN KEY (IDAsignatura)
 REFERENCES ASIGNATURA (AsignaturaID);
```

ANEXO 2: Consultas SQL.

- ***Por fecha de matriculación, titulación y alumno obtener el número total de créditos troncales superados.***

```
SELECT SUM (cred_tronc_supera_1er_curso
            +cred_tronc_supera_2_curso
            +cred_tronc_superados_3er_curso
            +cred_tronc_superados_4_curso
            +cred_tronc_supera_5_curso
            +cred_tronc_superados_6_curso) CREDITOS
FROM   Seguim_Academico h, titulacion tit, alumno al
   WHERE   Fec_matriculacion_1er_curso > '01-09-2004'
   AND     h.IDAlumno = al.IDAlumno
   AND     h.IDTitulacion =tit.IDTitulacion
   AND     al.IDAlumno =1
GROUP BY h.idtitulacion, h.IDAlumno;
```

- ***Por fecha de matriculación, titulación y alumno obtener el número total de créditos troncales pendientes de superar.***

```
SELECT SUM (cred_tronc_pendientes_1er_cur
            + cred_tronc_pendientes_2_cur
            + cred_tronc_pendientes_3er_cur
            + cred_tronc_pendientes_4_cur
            + cred_tronc_pendientes_5_curso
            + cred_tronc_pendientes_6_curso) CREDITOS
FROM   Seguim_Academico h, titulacion tit, alumno al
   WHERE   Fec_matriculacion_1er_curso > '01-09-2004'
   AND     h.IDAlumno = al.IDAlumno
   AND     h.IDTitulacion =tit.IDTitulacion
   AND     al.IDAlumno =1
GROUP BY  h.idtitulacion, h.IDAlumno;
```

- ***Número de alumnos con fecha de aprobado del primer curso mayor en un año a la fecha de matriculación de primer curso, agrupados por titulación.***

```
SELECT count (*)
FROM alumno al, titulacion tit, Seguim_Academico h
   WHERE h.Fec_aprob_complet_1er_curso-h.Fec_matriculacion_1er_curso>365
GROUP BY h.idtitulacion;
```

- ***Número de alumnos con fecha de aprobado del segundo curso mayor en un año a la fecha de matriculación de primer curso, agrupados por titulación.***

```
SELECT count (*)
FROM alumno al, titulacion tit, Seguim_Academico h
   WHERE  h.Fec_aprob_complet_2_curso - h.Fec_matriculacion_2_curso > 365
GROUP BY h.idtitulacion;
```

- ***Número de alumnos con fecha de aprobado del tercer curso mayor en un año a la fecha de matriculación de primer curso, agrupados por titulación.***

```
SELECT count (*)
FROM alumno al, titulacion tit, Seguim_Academico h
WHERE  h.Fec_aprob_complet_3er_curso – h.Fec_matriculacion_3er_curso > 365
GROUP BY h.idtitulacion;
```

- ***Número de alumnos con fecha de aprobado del cuarto curso mayor en un año a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay cuarto curso)***

```
SELECT count (*)
FROM alumno al, titulacion tit, Seguim_Academico h
WHERE  h.Fec_aprob_complet_4_curso - h.Fec_matriculacion_4_curso > 365
GROUP BY h.idtitulacion;
```

- ***Número de alumnos con fecha de aprobado del quinto curso mayor en un año a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay quinto curso)***

```
SELECT count (*)
FROM alumno al, titulacion tit, Seguim_Academico h
WHERE  h.Fec_aprob_complet_5_curso - h.Fec_matriculacion_5_curso > 365
GROUP BY h.idtitulacion;
```

- ***Número de alumnos con fecha de aprobado del sexto curso mayor en un año a la fecha de matriculación de cuarto curso agrupados por titulación. (si hay sexto curso)***

```
SELECT count (*)
FROM alumno al, titulacion tit, Seguim_Academico h
WHERE  h.Fec_aprob_complet_6_curso - h.Fec_matriculacion_6_curso > 365
GROUP BY h.idtitulacion;
```

- ***Titulación con nota media de primer curso más alta.***

```
SELECT  h.Nota_media_primer_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
          (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_primer_curso DESC;
```

- ***Titulación con nota media de primer curso más baja.***

```
SELECT  h.Nota_media_primer_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
           (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_primer_curso ASC;
```

- ***Titulación con nota media de segundo curso más alta.***

```
SELECT  h.Nota_media_segundo_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
           (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_segundo_curso DESC;
```

- ***Titulación con nota media de segundo curso más baja.***

```
SELECT  h.Nota_media_segundo_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
           (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_segundo_curso ASC;
```

- ***Titulación con nota media de tercer curso más alta.***

```
SELECT  h.Nota_media_tercer_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
           (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_tercer_curso DESC;
```

- ***Titulación con nota media de tercer curso más baja.***

```
SELECT  h.Nota_media_tercer_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
           (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_tercer_curso ASC;
```

- ***Titulación con nota media de cuarto curso más alta.***

```
SELECT  h.Nota_media_cuarto_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
          (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_cuarto_curso DESC;
```

- ***Titulación con nota media de cuarto curso más baja.***

```
SELECT  h.Nota_media_cuarto_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
          (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_cuarto_curso ASC;
```

- ***Titulación con nota media de quinto curso más alta.***

```
SELECT  h.Nota_media_quinto_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
          (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_quinto_curso DESC;
```

- ***Titulación con nota media de quinto curso más baja.***

```
SELECT  h.Nota_media_quinto_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
          (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_quinto_curso ASC;
```

- ***Titulación con nota media de sexto curso más alta.***

```
SELECT  h.Nota_media_sexto_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
     AND tit.IDTitulacion IN
          (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
     AND rownum <=1
ORDER BY Nota_media_sexto_curso DESC;
```

```
SELECT  h.Nota_media_sexto_curso, h.IDTitulacion
FROM    titulacion tit, Seguim_Academico h
  WHERE h.IDTitulacion =tit.IDTitulacion
    AND tit.IDTitulacion IN
        (SELECT DISTINCT IDtitulacion FROM Seguim_Academico)
    AND rownum <=1
ORDER BY Nota_media_sexto_curso ASC;
```

- ***Alumnos con la mayor diferencia entre la fecha de matriculación y la fecha de finalización de la carrera, agrupados por titulación.***

```
SELECT  MAX (Fecha),  IDAlumno, IDTitulacion FROM
 (SELECT  h.IDAlumno, tit.IDTitulacion,
        (h.Fecha_terminacion_titulacion – h.Fec_matriculacion_1er_curso)Fecha
        FROM alumno al, titulacion tit, Seguim_Academico h
        WHERE al.IDAlumno = h.IDAlumno
        AND   tit.IDTitulacion = h.IDTitulacion
        AND   tit.IDTitulacion in
              (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la mayor diferencia entre la fecha de finalización de la carrera y la fecha de finalización del Proyecto Fin de Carrera, agrupados por titulación.***

```
SELECT  MAX (Fecha),  IDAlumno, IDTitulacion FROM
 (SELECT  h.IDAlumno, tit.IDTitulacion,
        (h.Fecha_realizacion_PFC – h.Fecha_terminacion_titulacion)Fecha
        FROM alumno al, titulacion tit, Seguim_Academico h
        WHERE al.IDAlumno = h.IDAlumno
        AND   tit.IDTitulacion = h.IDTitulacion
        AND   tit.IDTitulacion in
              (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más alta de las asignaturas troncales superadas en primer curso agrupada por titulaciones.***

```
SELECT  MAX (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
             media_asig_tronc_supera_1o Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
        AND tit.IDTitulacion = h.IDTitulacion
        AND tit.IDTitulacion in
            (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más baja de las asignaturas troncales superadas en primer curso agrupada por titulaciones.***

```
SELECT  MIN (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                media_asig_tronc_supera_1o Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
        AND tit.IDTitulacion = h.IDTitulacion
        AND tit.IDTitulacion in
            (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más alta de las asignaturas obligatorias superadas en primer curso agrupada por titulaciones.***

```
SELECT  MAX (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                media_asig_oblig_supera_1o Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
        AND tit.IDTitulacion = h.IDTitulacion
        AND tit.IDTitulacion in
            (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más baja de las asignaturas obligatorias superadas en primer curso agrupada por titulaciones.***

```
SELECT  MIN (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                media_asig_oblig_supera_1o Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
        AND tit.IDTitulacion = h.IDTitulacion
        AND tit.IDTitulacion in
            (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más alta de las asignaturas optativas superadas en primer curso, agrupados por titulaciones.***

```
SELECT  MAX (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                media_asig_optat_supera_1o Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
        AND tit.IDTitulacion = h.IDTitulacion
        AND tit.IDTitulacion in
            (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más baja de las asignaturas optativas superadas en primer curso agrupada por titulaciones.***

```
SELECT  MIN (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                media_asig_optat_supera_1o Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
          AND tit.IDTitulacion = h.IDTitulacion
          AND tit.IDTitulacion in
                (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más alta en la carrera, agrupados por titulación***.

```
SELECT  MAX (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                Nota_media_carrera Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
          AND tit.IDTitulacion = h.IDTitulacion
          AND tit.IDTitulacion in
                (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

- ***Alumnos con la nota media más baja en la carrera, agrupados por titulación***.

```
SELECT  MIN (Nota), IDAlumno, IDTitulacion FROM
    (SELECT  h.IDAlumno, tit.IDTitulacion,
                Nota_media_carrera Nota
       FROM alumno al, titulacion tit, Seguim_Academico h
      WHERE al.IDAlumno = h.IDAlumno
          AND tit.IDTitulacion = h.IDTitulacion
          AND tit.IDTitulacion in
                (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```